Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 46

5-1-2017

Vol. 16, No. 1 (Full Issue)

JMASM Editors

Vol. 16, No. 1 (Full Issue)

Erratum

Montez-Rath, M. E., Kapphahn, K., Mathur, M. B., Mitani, A. A., Hendry, D. J., & Desai, M. (2017). Guidelines for generating right-censored outcomes from a Cox model extended to accommodate time-varying covariates. Journal of Modern Applied Statistical Methods, 16(1), 86-106. doi: 10.22237/jmasm/1493597100

The initial publication indicated that "studies" was the subject of the first sentence of the abstract. The subject is more properly "simulating," and the verb "is" is conjugated accordingly.

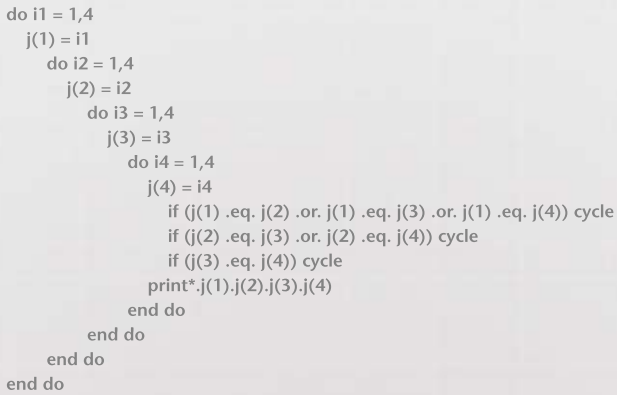
Beauducel, A. (2017). A Schmid-Leiman-based transformation resulting in perfect inter-correlations of three types of factor score predictors. Journal of Modern Applied Statistical Methods, 16(1), 107-126. doi: 10.22237/jmasm/1493597160

In the initial publication of this paper, the editors omitted the final term, $\text{diag}(\Lambda'\Sigma^{-1}\Lambda)^{1/2}$, from the first line of Equation 15 on page 114, placing it in the second line erroneously. This inversion has been corrected.

Takahashi, M. (2017). Multiple ratio imputation by the EMB algorithm: theory and simulation. Journal of Modern Applied Statistical Methods, 16(1), 630-656. doi: 10.22237/jmasm/1493598840

While this article appears in 16(1) in the section Algorithms and Code, it was submitted and accepted to this issue's Regular Articles, and the placement is an oversight of the Editors. For reasons of pagination, reassignment of the article is problematic; however, it received a full double-blind peer review before acceptance, as do all Regular Articles in *JMASM*.

בס"ד



Journal of Modern Applied Statistical Methods

Invited Articles
Rand Wilcox, Chauncy M. Dayton,
and Vance Berger

Vol. 16, No. 1 • May, 2017

בס"ד
Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, ii-vii
doi: 10.22237/jmasm/1493599440

ISSN 1538 − 9472

Journal of **Modern Applied Statistical Methods**

JMASM (ISSN 1538−9472, http://digitalcommons.wayne.edu/jmasm) is an independent, open access electronic journal, published biannually in May and November by JMASM Inc. (PO Box 48023, Oak Park, MI, 48237) in collaboration with the Wayne State University Library System. *JMASM* seeks to publish (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo- random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal correspondence (other than manuscript submissions) and requests for advertising may be forwarded to ea@jmasm.com. See back matter for instructions for authors.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, ii-vii
doi: 10.22237/jmasm/1493599440

ISSN 1538 − 9472

Journal of Modern Applied Statistical Methods

Vol. 16, No. 1

May 2017

Table of Contents

Invited Articles

Regular Articles

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, ii-vii
doi: 10.22237/jmasm/1493599440

ISSN 1538 − 9472

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, ii-vii
doi: 10.22237/jmasm/1493599440

ISSN 1538 − 9472

Emerging Scholars

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, ii-vii
doi: 10.22237/jmasm/1493599440

ISSN 1538 − 9472

Algorithms and Code

Statistical Software Applications & Review

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, ii-vii
doi: 10.22237/jmasm/1493599440

ISSN 1538 − 9472

Book Review

Letters To The Editor

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 2.

ISSN 1538 – 9472

Invited Articles

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 3-19.
doi: 10.22237/jmasm/1493596800

ISSN 1538 − 9472

Robust ANCOVA: Confidence Intervals That Have Some Specified Simultaneous Probability Coverage When There Is Curvature And Two Covariates

Rand Wilcox
University of Southern California
Los Angeles, California

Consider the commonly occurring situation where the goal is to compare two independent groups and there are two covariates. Let $M_j(X)$ be some conditional measure of location for the j^{th} group associated with some random variable Y given $X = (X_1, X_2)$. The goal is to H_0: $M_1(X) = M_2(X)$ for each $X \in \Omega$ in a manner that controls the probability of one or more Type I errors. An extant technique (method M_1 here) addresses this goal without making any parametric assumption about $M_j(X)$. However, a practical concern is that it does not provide enough detail regarding where the regression surfaces differ, due to using a very small number of covariate points, which can result in relatively low power. Method M_2 was proposed for testing the global hypothesis H_0: $M_1(X) = M_2(X)$ for all $X \in \Omega$, which offers a distinct power advantage over method M_1. It uses the deepest half of the covariate points rather than small number of points used by method M_1. However, method M_2 does not provide any details about which covariate points yield a significant result. A multiple comparison procedure is proposed that deals with this shortcoming of method M_2, and simultaneously it can provide higher power than method M_1.

Keywords: ANCOVA, trimmed mean, smoothers, Well Elderly 2 study

Introduction

Consider the common situation where the goal is to compare two independent groups based on two covariates. The classic ANCOVA (analysis of covariance) method assumes that

$$Y_j = \beta_{0j} + \beta_1 X_{1j} + \beta_2 X_{2j} + \varepsilon, \quad (1)$$

Rand Wilcox is an Professor in the Department of Psychology. Email him at: rwilcox@usc.edu.

where β_{0j}, β_1 and β_2 are unknown parameters estimated via least squares regression and ε is a random variable having a normal distribution with mean zero and unknown variance σ^2. So the regression planes are assumed to be parallel and the groups can be compared by testing

$$H_0 : \beta_{01} = \beta_{02}, \tag{2}$$

the hypothesis that the intercepts are equal. It is well known, however, that least squares regression is not robust (e.g., Staudte and Sheather, 1990; Maronna et al., 2006; Heritier et al., 2007; Hampel et al., 1986; Huber and Ronchetti, 2009; Wilcox, 2012). A practical consequence is that power can be relatively low even under a small departure from normality. Moreover, even a single outlier can yield a poor fit to the bulk of the points when using least squares regression.

Another concern with the classic ANCOVA model is that two types of homoscedasticity are assumed. The first is that for each group, the variance of the error term does not depend on the value of the covariate. If this assumption is violated the wrong standard error is being used (e.g., Long & Ervin, 2000). A seemingly natural way of justifying a homoscedastic error term is to test the assumption that it is indeed homoscedastic. However, Ng and Wilcox (2011) found that this strategy is unsatisfactory. The problem is that methods for testing the homoscedasticity assumption do not have enough power to detect situations where heteroscedasticity is a practical concern. The second homoscedasticity assumption is that the variance of the error term is the same for both groups. Violating these assumptions can result in poor control over the Type I error probability.

Yet another fundamental concern with (1) is that the true regression surfaces are assumed to be planes. Presumably this is a reasonable approximation in some situations, but experience with smoothers (e.g., Hastie & Tibsherani, 1990; Wilcox, 2012) made it clear that often this is not the case. When there is curvature, using some obvious parametric regression model might suffice. (For example, include a quadratic term.) It is known that this approach can be inadequate, which has led to a substantial collection of nonparametric regression methods, often called smoothers, for dealing with curvature in a more flexible manner (e.g., Härdle, 1990; Efromovich, 1999; Eubank, 1999; Fox, 2001; Györfi, et al., 2002).

One more limitation of the classic model is the assumption that the regression surfaces are parallel. The assumption that the slope parameters are equal could be tested, but it is unclear when such a test has enough power to

detect situations where this assumption is violated to the point that it makes a practical difference.

Let $M_j(X)$ be some conditional measure of location associated with Y given $X = (X_1, X_2)$, where $M_j(X)$ is some unknown function. Here, the model given by (1) is replaced with the less restrictive model

$$Y_j = M_j(X) + \lambda(X)\varepsilon_j, \tag{3}$$

where $\lambda(X)$ is some unknown function used to model heteroscedasticity. The random variable ε_j has some unknown distribution with variance σ_j^2. So unlike the classic approach where it is assumed that

$$M_j(X) = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2,$$

no parametric model for $M_j(X)$ is specified and $\sigma_1 = \sigma_2$ is not assumed. In particular it is not assumed that the regression surfaces are parallel.

Let $X_1, \ldots, X_K$ be K covariate points that are chosen empirically in a manner to be described. The goal here is to test the K hypotheses

$$H_0 : M_1(X_k) = M_2(X_k) \tag{4}$$

for each $k = 1, \ldots, K$ such that the probability of one or more Type I errors is approximately equal to α. The focus is on situations where $M_j(X)$ is a trimmed mean, but the basic strategy underlying the proposed approach (method M_3 in the so-named section) can in principle be extended to other robust measures of location.

Wilcox (2012) suggested a simple method for testing (4) for each $k = 1, \ldots, K$ when the covariate points are chosen based on how deeply they are nested within the cloud of covariate points (this is method M_1 in the so-named section). The K points are chosen to include the point in the first group having the deepest half space depth plus the points on the .5 depth contour. This typically results in using a fairly small number of covariate points where the corresponding Y values are compared based on a robust measure of location. Among the K tests that are performed, the probability of one or more Type I errors can be controlled using some improvement on the Bonferroni method (e.g., Hommel, 1988; Hochberg, 1988). However, it is not clear when this relatively simple approach will choose covariate values that are likely to detect true differences between the

groups. Another concern is that important details about where the groups differ will be missed due to using a small number of covariate points.

A way of dealing with this issue is to select a larger collection of covariate points. The strategy here is to use the deepest half of the covariate points in the first group. But as K increases, an obvious concern is the negative impact this will have on power when using the methods derived by Hommel (1988) and Hochberg (1988). A method that controls the false discovery rate when dealing with dependent test statistics (e.g., Benjamini &Yekutieli, 2001) suffers from the same concern. Wilcox (2016) derived a method for testing the global hypothesis that (4) is true for the deepest half of the covariate points in the first group (this is method M_2 in the so-named section). However, when this method rejects, it provides virtually no information about which of the individual hypotheses can be rejected.

The goal here is to suggest a method for controlling the probability of one or more Type I errors when testing the K hypotheses given by (4). Like method M_2, the deepest half of the covariate points is used. But rather than use the methods derived by Hommel (1988) and Hochberg (1988), an alternative technique is suggested that has a certain similarity to using a Studentized maximum modulus distribution.

Description of the Methods

Let (Y_{ij},X_{ij}) $(i = 1, ..., n_j; j = 1, 2)$ be a random sample from the j^{th} group. The methods compared here are based in part on a method derived by Yuen (1974) for comparing the population trimmed means of two independent groups. To describe it, momentarily ignore the covariates and consider the goal of testing

$$H_0 : \mu_{t1} = \mu_{t2}, \tag{5}$$

the hypothesis that two independent groups have equal population trimmed means. For the j^{th} group $(j = 1, 2)$, let $Y_{(1)j} \leq \ldots \leq Y_{(n_j)j}$ denote the Y_{ij} values written in ascending order. For some $0 \leq \gamma < .5$, the γ-trimmed mean for the j^{th} group is

$$\bar{Y}_j = \frac{1}{n_j - 2g_j}\left(Y_{(g_j+1)j} + \cdots + Y_{(n_j-g_j)j}\right)$$

where $g_j = [\gamma n_j]$ is the greatest integer less than or equal to γn_j. Here the focus is on $\gamma = .2$, a 20% trimmed mean. Under normality, this choice has good efficiency

relative to the sample mean (Rosenberger & Gakso, 1983). Moreover, the sample 20% trimmed mean enjoys certain theoretical advantages. First, it has a reasonably high breakdown point, which refers to the proportion of values that must be altered to destroy it. Asymptotic results and simulations indicate that it reduces substantially concerns about the impact of skewed distributions on the probability of a Type I error (e.g., Wilcox, 2012). This is not to suggest that 20% trimming is always the optimal choice: clearly this is not the case. The only suggestion is that it is a reasonable choice among the many robust estimators that might be used.

Winsorizing the Y_{ij} values refers to setting

$$W_{ij} = Y_{(g_j+1)}, \text{ if } Y_{ij} \le Y_{(g_j+1)}$$
$$W_{ij} = Y_{ij}, \text{ if } Y_{(g_j+1)} < Y_{ij} < Y_{(n_j-g_j)}$$
$$W_{ij} = Y_{(n_j-g_j)}, \text{ if } Y_{ij} \ge Y_{(n_j-g_j)}$$

The Winsorized sample mean corresponding to group j is the mean based on the Winsorized values, and the Winsorized variance, s_{wj}^2, is the usual sample variance, again based on the Winsorized values.

Let $h_j = n_j - 2g_j$. That is, h_j is the number of observations left in the j^{th} group after trimming. Let

$$d_j = \frac{\left(n_j - 1\right)s_{wj}^2}{h_j\left(h_j - 1\right)}.$$

Yuen's test statistic is

$$T_y = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{d_1 + d_2}}$$

The null distribution is taken to be a Student's t distribution with degrees of freedom

$$\hat{\upsilon} = \frac{\left(d_1 + d_2\right)^2}{C},$$

where

$$C = \frac{d_1^2}{h_1} + \frac{d_2^2}{h_2}$$

Method M_1

Method M_1 was described in Wilcox (2012, section 11.11.3). A complete description of the many computational details is not provided here, but an outline of the method is provided with the goal of explaining how it differs from methods M_2 and M_3.

Momentarily consider a single covariate point, X. For fixed j, method M_1 estimates $M_j(X)$ using the Y_{ij} associated with the X_{ij} points that are close to X. More precisely, for the jth group, compute a robust covariance matrix based on $X_{ij}(i = 1, \ldots, n_j)$. There are many ways of computing a robust covariance matrix with no single estimator dominating. Here a skipped covariance matrix is used, which is computed as follows. For fixed j, outliers among the X_{ij} points are identified using a projection-type multivariate outlier detection technique (e.g., Wilcox, 2012, section 6.4.9). These outliers are removed and the usual covariance matrix is computed using the remaining data.

Next, compute robust Mahalanobis distances for each covariate point based on the robust covariance matrix just described, with X taken to be the center of the data. The point X_{ij} is said to be close to X if its robust Mahalanobis distance is small, say less than or equal to f, which is called the span. Generally, $f = .8$ performs reasonably well when the goal is to approximate the regression surface. Of course exceptions are encountered, but henceforth $f = .8$ is assumed. Let $P_j(X)$ be the subset of $\{1, 2, ..., n_j\}$ that indexes the X_{ij} values such that the Mahalanobis distance associated with X_{ij} is less than or equal to f. Let $N_j(X)$ be the cardinality of the set $P_j(X)$ and let $M_j(X)$ denote the 20% trimmed mean based on the Y_{ij} values for which $i \in P_j(X)$. Then for the single point X, (4) can be tested by applying Yuen's method with the Y_{ij} values for which $i \in P_j(X)$ provided both $N_1(X)$ and $N_2(X)$ are not too small. Following Wilcox (2012), this is taken to mean that Yuen's method can be applied if simultaneously $N_1(X) \geq 12$ and $N_2(X) \geq 12$, in which case the two groups are said to be comparable at X.

Consider the issue of choosing covariate points where the regression surfaces will be compared. For the first group, compute how deeply each X_{i1} is nested within the cloud of covariate points ($i = 1, \ldots, n_j$). This is done with a projection-type method that is similar to an approach discussed by Donoho and

Gasko (1992). The many computational details are not described and are not particularly important for present purposes. Here it is merely noted that an approximation of halfspace depth is used, which is described in Wilcox (2012, section 6.2.3) and labeled approximation A_1. Consider the deepest point as well as those on the polygon containing the central half of the data. (Liu et al., 1999, call this polygon the .5 depth contour.) Method M_1 applies Yuen's method at each of these points provided the regression surfaces are comparable at these points as previously defined. The probability of one or more Type I errors is controlled using the method in Hochberg (1988).

Method M_2

There are several positive features of method M_1 but some negative features as well. First, Yuen's method for comparing trimmed means has been studied extensively and appears to perform relatively well in terms of both Type I errors and power. The method for choosing the covariate values seems reasonable in the sense that it uses points that are nested deeply within the cloud of covariate points, which reflect situations where the regression surfaces are comparable. Roughly, deeply nested points correspond to situations where the regression surfaces can be estimated in a relatively accurate manner. If a point X is not deeply nested in the cloud of covariate values, finding a sufficiently large number of other points that are close to X might be impossible.

But a concern with M_1 is that perhaps true differences might be missed because typically a relatively small number of covariate points are used. In the Illustration to follow, only three covariate points are used by M_1, with sample sizes 187 and 228. Method M_2 deals with this concern in the following manner. First, it computes the projection depth for each X_{i1} (the i^{th} covariate vector in group 1) in the same manner as method M_1. Let the set $\{X_1, \ldots, X_K\}$ indicate the deepest half of the points in the first group. Points where the regression surfaces are not comparable (i.e., $N_1(X) < 12$ or $N_2(X) < 12$) are discarded. Because K can be relatively large, it is approximately equal to $n_1/2$, controlling the probability of one or more Type I errors via Hochberg's method or Hommel's method is likely to have relatively low power.

The reason for choosing the deepest half of the covariate points, rather than some larger proportion, is that typically the regression surfaces are comparable at all K points when the sample sizes for both groups are greater than or equal to 50. For a larger proportion of points, this is often not the case. There are, of course,

many other variations. Some other measure of depth might be used or one could use all of the covariate points where the regression surfaces are comparable.

Method M_2 proceeds in the same manner as method M_1 by testing H_0: $M_1(X) = M_2(X)$ for each $X \in \{X_1, \ldots, X_K\}$. Label the resulting p-values $p_1, \ldots, p_K$. The idea is to test the global hypothesis that (4) is true for every $k = 1, \ldots, K$ using some function of these K p-values. Perhaps the best-known method for testing some global hypothesis based on p-values is a technique derived by Fisher (1932). But Zaykin et al. (2002) note that the ordinary Fisher product test loses power in cases where there are a few large p-values. They suggest using instead a truncated product method (TPM), which is based on the test statistic

$$W = \prod p_k^{I(p_k \leq \tau)}$$

where I is the indicator function (cf. Li & Siegmund, 2015). Setting $\tau = 1$ yields Fisher's method, but Zaykin et al. suggest using $\tau = .05$. Zaykin et al. derive the null distribution of W when all K tests are independent. But the K tests performed here are not independent simply because $P_j(X_k) \cap P_j(X_l)$, $k \neq l$, is not necessarily empty. If this dependence among the tests is ignored when computing a critical value for W, control over the Type I error probability is poor. For the dependent case, Zaykin et al. suggest using a bootstrap method, but this results in relatively high execution time for the situation at hand making this approach difficult to study via simulations. Consequently, an alternative approach was used: Momentarily assume normality and homoscedasticity with the goal of determining the α quantile of W, say w, in which case (4) is rejected at the α level if $W \leq w$. Then study the impact of non-normality and heteroscedasticity via simulations.

The critical value w is determined via simulations using (2) with $M_j(X) \equiv 0$ and ε_j having a standard normal distribution. More precisely, for each j, (Y_{ij}, X_{ij}) $(i = 1, \ldots, n_j; j = 1,2)$ are generated from a trivariate normal distribution where all correlations are zero. Then W is computed and this process is repeated say B times yielding $W_1, \ldots, W_B$. Put these B values in ascending order yielding $W_{(1)} \leq \ldots \leq W_{(B)}$. Then w is estimated to be $W_{(k)}$, where k is αB rounded to the nearest integer. Here, $B = 4000$ is used. Increasing the correlation to .5 had almost no impact on the estimated critical value.

One of many alternative methods is to use instead the test statistic

$$\bar{Q} = \frac{1}{K}\sum p_k$$

Wilcox (2016) found that this alternative test statistic performed relatively well, in terms of power, under a shift in location model. Now reject the global hypothesis if $\bar{q} \leq q_\alpha$, the α quantile of $\bar{Q}$, which again is determined via simulations in the same manner as the critical value w. So rejecting indicates that one or more of the hypotheses given by (4) are false, but details about which ones are lacking.

Method M_3

The following strategy, called method M_3, is suggested for dealing with the limitation of method M_2. First, choose covariate points as done by method M_2. Based on this process for choosing covariate points, determine p_α, the α quantile of the distribution of the minimum p-value returned by method M_2. This is done via simulations in essentially the same manner used by method M_2. The only difference from method M_2 is that W and $\bar{Q}$ are replaced by $\tilde{p} = \min(p_1, ..., p_K)$. So for a simulation based on B replications yielding $\tilde{p}_1, \ldots, \tilde{p}_B$, p_α is estimated with $\tilde{p}_{(k)}$, where k is the same as in method M_2 and $\tilde{p}_{(1)} \leq \cdots \leq \tilde{p}_{(B)}$ are the $\tilde{p}$ values written in ascending order. Then make a decision about whether $M_1(X)$ is larger than $M_2(X)$ for any covariate point for which the corresponding p-value is less than or equal to p_α. Otherwise, no decision is made. So method M_3 has the potential of providing more detail about where the regression surfaces differ. But of course there is the issue of how well it performs when dealing with non-normality, heteroscedasticity and curvature, which is examined via simulations in the next section. And another issue is the impact on power compared to method M_1.

Table 1. Some estimates of p_α, $\alpha = .05$

n	p_α	n	p_α	n	p_α
50	0.00458	80	0.00248	400	0.00131
55	0.00320	100	0.00186	500	0.00135
60	0.00282	200	0.00142	600	0.00108
70	0.00259	300	0.00142	800	0.00096

Estimates of p_α, when $n_1 = n_2 = n$ are informative. Table 1 shows estimates for values of n ranging between 50 and 800 when $\alpha = .05$. So the estimates appear

to be converging to zero, but at an extremely slow rate. Consider, for example $n = 100$, in which case fifty hypotheses are tested. As indicated by Table 1, p_α is estimated to be .00186. Using the Bonferroni method instead, each hypothesis would be tested at the .0005 level, which is even less than the estimate of pα when using M_3 with n = 800.

Simulation Results

As is evident, an issue is the impact on the Type I error probability when dealing with non-normal distributions as well as situations where there is an association with the covariate variables. Simulations were used to address this issue with $n_1 = n_2 = 50$. Smaller sample sizes, such as $n_1 = n_2 = 30$, routinely result in situations where no covariate values can be found where comparisons can be made. That is, $N_1(X) < 12$ or $N_2(X) < 12$ for all $X \in \{X_1, \ldots, X_K\}$.

Estimated Type I error probabilities were based on 4000 replications. Four types of distributions were used: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, values for the error term ε_j in (3) were generated from one of four g-and-h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. If Z has a standard normal distribution, then by definition

$$V = \frac{\exp(gZ) - 1}{g} \exp\left(hZ^2/2\right), \text{ if } g > 0$$
$$V = Z \exp\left(hZ^2/2\right), \text{ if } g = 0$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 2 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). The X_{ij} values were generated from a bivariate normal distribution with correlation equal to zero. Increasing this correlation to .5 altered the estimates of the Type I error probability by only a few units in third decimal place, so for brevity they are not reported.

Table 2. Some properties of the *g*-and-*h* distribution.

g	*h*	κ_1	κ_2
0.00	0.00	0.00	3.00
0.00	0.20	0.00	21.46
0.20	0.00	0.61	3.68
0.20	0.20	2.81	155.98

Two types of regression surfaces were considered. The first deals with the situation where $Y = \lambda(X)\varepsilon$, which is labeled S_1. The second, labeled S_2, is $Y = X^2 + \lambda(X)\varepsilon$. Three choices for $\lambda(X)$ were considered: $\lambda(X_i) \equiv 1$ (VP_1), $\lambda(X_i) = |X_{i1}| + 1$ (VP_2) and $\lambda(X_i) = 1/(|X_{i1}| + 1)$ (VP_3). Estimated Type I error probabilities are reported in Table 3. Although the seriousness of a Type I error depends on the situation, Bradley (1978) suggested as a general guide, when testing at the .05 level, the actual level should be between .025 and .075. He goes on to suggest that ideally the actual level should be between .045 and .055. As can be seen, the estimates satisfy his first criterion, and nearly all of them satisfy his more stringent criterion.

Table 3. Estimated Type I error probabilities when testing at the $\alpha = .05$ level, $n_1 = n_2 = 50$

g	*h*	*S*	VP_1	VP_2	VP_3
0.000	0.000	1.000	0.050	0.052	0.050
0.000	0.000	2.000	0.056	0.050	0.048
0.000	0.200	1.000	0.046	0.039	0.049
0.000	0.200	2.000	0.048	0.050	0.053
0.200	0.000	1.000	0.052	0.050	0.044
0.200	0.000	2.000	0.054	0.048	0.050
0.200	0.200	1.000	0.051	0.048	0.048
0.200	0.200	2.000	0.055	0.040	0.044

In some situations, method M_2 can have substantially higher power than method M_3, where power is taken to be the probability of detecting one or more true differences. Consider, for example, the situation where for the first group $Y = \varepsilon$ and for the second group $Y = \varepsilon + .5$, where ε has a standard normal distribution and both sample sizes are 50. Then method M_2 has power approximately .41 compared to .26 using method M_3. If instead $Y = X^2 + \varepsilon$ for the second group, now power is .79 for M_2 and .65 using M_3. That is, M_2 might offer a substantial gain in power among the situations considered here at the expense of

providing virtually no details about where significant results are obtained. However, methods M_2 and M_3 are sensitive to different features among the p-values. The next section illustrates that situations are encountered where M_3 rejects in contrast to M_2.

To provide some sense of how methods M_3 and M_1 compare in terms of power, again consider the situation where for the first group $Y = \varepsilon$ and for the second group $Y = \varepsilon + .5$. With both sample sizes equal to 100, power was estimated to be .51 and .42 for M_3 and M_1, respectively. If instead $Y = X^2 + \varepsilon$ for the second group, now power is .52 for M_3 and .51 using M_1. If $Y = X + \varepsilon$ for the second group, now the corresponding power estimates are .62 and .55. So, for at least some situations, method M_3 has substantially higher power than method M_1 despite the substantially larger number of hypotheses that are tested.

Illustrations

Data from the Well Elderly 2 study (Clark et al., 2011; Jackson et al., 2009) are used to illustrate that the choice between M_2 and M_3 can make a practical difference. A general goal in the Well Elderly 2 study was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study was aimed at understanding the impact of intervention on a measure of perceived physical health, which was measured with the RAND 36-item (SF36) Health Survey, a measure of self-perceived physical health and mental well-being (Hays et al., 1993; McHorney et al., 1993). Higher scores reflect greater perceived health and well-being. There were two covariates. The first is a measure of depressive symptoms based on the Center for Epidemiologic Studies Depressive Scale (CESD). The CESD (Radloff, 1977) is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Lewinsohn et al., 1988; Foley et al., 2002). Higher scores indicate a higher level of depressive symptoms. The other covariate was the cortisol awakening response (CAR), which is defined as the change in cortisol concentration that occurs during the first hour after waking from sleep. Extant studies (e.g., Clow et al., 2004; Chida & Steptoe, 2009) indicated that measures of stress are associated with the CAR. (The CAR is taken to be the cortisol level upon awakening minus the level of cortisol after the participants were awake for about an hour.) The sample size for the control group was 187 and the sample size for the group that received intervention was 228.

Based on both methods M_1 and M_2, no significant differences were found when testing at the .05 level. Method M_1 used only three covariate points. In

contrast, method M_3 finds nine significant results among the 74 covariate points that were used. They occur where the CAR is negative (cortisol increases after awakening) and CESD is relatively low. So despite the simulation results indicating that M_2 can have higher power than M_3, situations are encountered where M_3 rejects and M_2 does not. Figure 1 shows a plot of the difference in SF36 scores (SF36 scores for the experimental group minus SF36 scores for the control group) as a function of the covariate points that were used. As can be seen, the largest differences occur when CESD scores are low and the CAR is negative. That is, intervention appears to be most beneficial, in terms of perceived health, for participants for whom cortisol increases after awakening. This is particularly true for participants who have low measures of depressive symptoms.

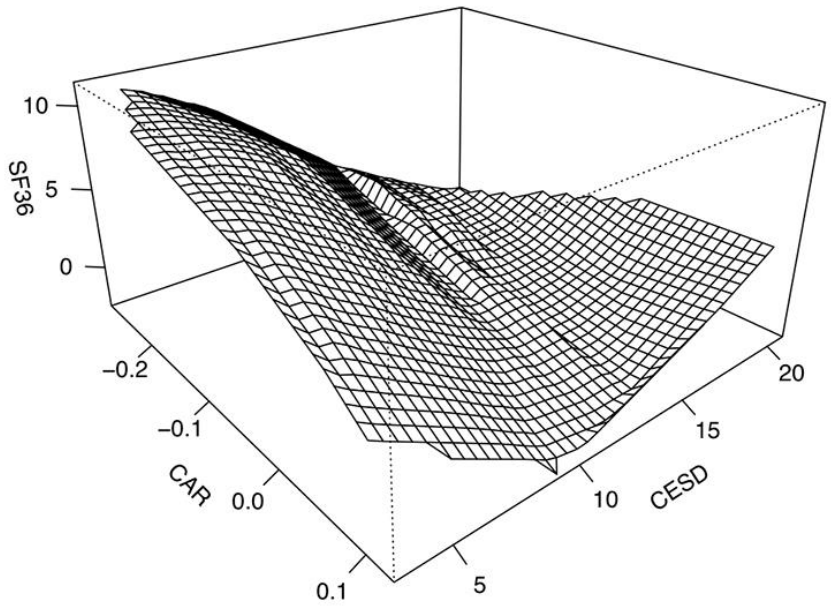


Figure 1. Regression surface predicting the typical difference in SF36 scores as a function of the CAR and CESD.

Conclusion

There are many variations of method M_3 that might have practical value. For example, some other measure of depth might be used or some alternative strategy for choosing the covariate points might offer an advantage. The main point is that

based on simulations, all indications are that method M_3 controls the probability of one or more Type I errors very well. At least in some situations it offers a distinct power advantage over M_1 and no situation has been found where the reverse is true. There are situations where M_2 provides higher power than M_3, but at the cost of providing almost no details about where a significant difference occurs among the covariate points that were used.

In principle, methods M_1, M_2 and M_3 can be used when there is more than two covariates. But a general concern is the curse of dimensionality: neighborhoods with a fixed number of points become less local as the dimensions increase (Bellman, 1961). In practical terms, the expectation is that as the number of covariates increases, it becomes increasingly difficult to get an accurate estimate of the true regression surface.

References

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton, NJ: Princeton University Press.

Benjamini Y. & Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165–1188. doi: 10.1214/aos/1013699998

Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Chida, Y. & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology*, *80*(3), 265–278. doi: 10.1016/j.biopsycho.2008.10.004

Clark, F., Jackson, J., Carlson, M., et al. (2011). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health*, *66*(9), 782–790. doi: 10.1136/jech.2009.099754

Clow, A., Thorn, L., Evans, P. & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress*, *7*(1), 29–37. doi: 10.1080/10253890410001667205

Donoho, D. L. & Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, *20*, 1803–1827. doi: 10.1214/aos/1176348890

Eakman, A. M., Carlson, M. E. & Clark, F. A. (2010). The meaningful activity participation assessment: a measure of engagement in personally valued activities. *International Journal of Aging Human Development, 70*(4), 299–317. doi: 10.2190/ag.70.4.b

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer-Verlag. doi: 10.1007/b97679

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.

Fisher, R. (1932). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Foley K., Reed P., Mutran E., et al. (2002). Measurement adequacy of the CESD among a sample of older African Americans. *Psychiatric Research, 109*(1), 61–69. doi: 10.1016/s0165-1781(01)00360-2

Fox, J. (2001). *Multiple and Generalized Nonparametric Regression*. Thousands Oaks, CA: Sage. doi: 10.4135/9781412985154

Györfi, L., Kohler, M., Krzyzk, A. & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer Verlag. doi: 10.1007/b97848

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.

Härdle, W. (1990). *Applied Nonparametric Regression*. Econometric Society Monographs No. 19, Cambridge, UK: Cambridge University Press.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Hays, R. D., Sherbourne, C .D. & Mazel, R. M. (1993). The Rand 36-item health survey 1.0. *Health Economics, 2*(3), 217–227. doi: 10.1002/hec.4730020305

Heritier, S., Cantoni, E, Copt, S. & Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. New York: Wiley. doi: 10.1002/9780470740538

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) *Exploring Data Tables Trends and Shapes*. New York: Wiley, pp. 461–515.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802. doi: 10.1093/biomet/75.4.800

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383–386. doi: 10.1093/biomet/75.2.383

Huber, P. J. & Ronchetti, E. (2009). *Robust Statistics* (2nd Ed). New York: Wiley. doi: 10.1002/9780470434697

Jackson, J., Mandel, D., Blanchard, J., et al. (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials, 6*(1), 90–101. doi: 10.1177/1740774508101191

Lewinsohn, P.M., Hoberman, H. M., Rosenbaum M. (1988). A prospective study of risk factors for unipolar depression. *Journal of Abnormal Psychology, 97*(3), 251–64. doi: 10.1037/0021-843x.97.3.251

Li, J. & Siegmund, D. (2015). Higher criticism: *p*-values and criticism. *Annals of Statistics, 43*(3), 1323–1350. doi: 10.1214/15-aos1312

Long, J. S. & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician, 54*(3), 217–224. doi: 10.1080/00031305.2000.10474549

Liu, R. Y., Parelius, J. M. & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics, 27*(3), 783–858. 10.1214/aos/1018031260

Maronna, R. A., Martin, D. R. & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. New York: Wiley. doi: 10.1002/0470010940

McHorney, C. A., Ware, J. E. & Raozek, A. E. (1993). The MOS 36-item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care, 31*(3), 247–263. doi: 10.1097/00005650-199303000-00006

Ng, M. & Wilcox, R. R. (2011). A comparison of two-stage procedures for testing least- squares coefficients under heteroscedasticity. *British Journal of Mathematical and Statistical Psychology, 64*(2), 244-258. doi: 10.1348/000711010x508683

Radloff, L. (1977). The CESD scale: a self report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401. doi: 10.1177/014662167700100306

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77*(3), 663–666. doi: 10.1093/biomet/77.3.663

Rosenberger, J. L. & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Understanding Robust and exploratory data analysis*. (pp. 297–336). New York: Wiley.

Staudte, R. G. & Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley. doi: 10.1002/9781118165485

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd Ed). San Diego, CA: Academic Press.

Wilcox, R. R. (2016). ANCOVA: A heteroscedastic global test when there is curvature and two covariates. *Computational Statistics, 31*(4), pp. 1593-1606. doi: 10.1007/s00180-015-0640-4

Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika, 61*(1), 165–170. doi: 10.1093/biomet/61.1.165

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology, 22*(2), 170–185. doi: 10.1002/gepi.0042

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 20-33.
doi: 10.22237/jmasm/1493596860

ISSN 1538 − 9472

A Reinterpretation and Extension of McNemar's Test

Chauncey M. Dayton
University of Maryland, College Park
College Park, MD

The McNemar test is extended to multiple groups based on a latent class model incorporating classes representing consistent responders and a single latent error rate. The method is illustrated with data from a CDC survey of immunizations for flu and pneumonia for which a part-heterogeneous model is selected for interpretation.

Keywords: McNemar test, latent class analysis, marginal homogeneity, response error

Introduction

The McNemar chi-square test is the procedure of choice in studies assessing marginal homogeneity for repeated dichotomous classifications. Typical applications involve two independent raters or assays providing dichotomous judgments for the same set of stimuli, or a panel of independent judges responding on two occasions to the same dichotomous variable. The research question is whether or not it is reasonable to describe the two marginal classification rates for, say, a positive classification as equivalent (i.e., homogeneous). The chi-square significance test for this case is attributed to McNemar (1947) and the generalization to square tables larger than 2×2 is often referred to as the Stuart-Maxwell test (Stuart, 1955; Maxwell, 1970). Although alternatives to the McNemar test have been proposed, the original procedure performs well in comparative simulations as shown by Fagerland, Lydersen, and Laake (2013). Also, methods for performing multiple comparisons involving several sets of 2×2 tables have been presented by Westfall, Troendle and Pennello (2010).

Dr. Dayton is Professor Emeritus of Measurement, Statistics and Evaluation, as well as principal researcher with BDS Data Analytics. Email him at: cdayton@umd.edu.

For dichotomous variables, *A* and *B*, let π_{ij} represent the theoretic proportion for level *i* of variable *A* and level *j* of variable *B* (Table 1). Marginal homogeneity implies that $\pi_{1.} = \pi_{.1}$ or

Table 1. Theoretic Proportions for 2 × 2 Table

	B+	B−	Row
A+	π_{11}	π_{12}	$\pi_{1.}$
A−	π_{21}	π_{22}	$\pi_{2.}$
Column	$\pi_{.1}$	$\pi_{.2}$	

equivalently, that $\pi_{2.} = \pi_{.2}$. Assuming a sample of *N* cases and observed frequencies, n_{ij}, this implies symmetry because $\pi_{1.} = N(\pi_{11} + \pi_{12})$ and $\pi_{.1} = N(\pi_{11} + \pi_{21})$ so that π_{12} must be equal to π_{21}. Note, however, that marginal homogeneity does not imply symmetry for tables larger than 3 × 3.

The test for symmetry and, per force, the test for marginal homogeneity, reduces to a two-celled goodness-of-fit test based on the observed frequencies n_{12} and n_{21} with the null hypothesis $\pi_{12} = \pi_{21}$, or equivalently, $\pi_{12} = \pi_{21} = .5$. Note that the expected frequencies are both equal to $(n_{12} + n_{21})/2$. In terms of observed frequencies, the McNemar statistic in the form of a Pearson chi-square, with one degree of freedom, can be written as: $\chi^2 = \dfrac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$.

An asymptotically equivalent test statistic can be based on a likelihood-ratio chi-square of the form $L^2 = -2\left(n_{12} Ln \dfrac{2n_{12}}{n_{12} + n_{21}} + n_{21} Ln \dfrac{2n_{21}}{n_{12} + n_{21}} \right)$. Often a correction for continuity is applied to the Pearson chi-square statistic to improve accuracy (Fleiss, 1981) and there are recent modifications such as mid-p computations (Fagerland, Lydersen & Laake, 2013). Agresti and Klingenburg (2005), and Klingenberg and Agresti (2006), have presented multivariate extensions of the McNemar test. Also, Durkalski, Palesch, Lipsitz, and Rust (2003) have introduced adaptations to account for clustering of observations.

The focus in the current study is on the issue of stratified homogeneity. Stratified homogeneity implies that marginal homogeneity for variables *A* and *B*, say, holds across the levels of a third variable (e.g., time, strata or groups). Feuer and Kessler (1989) considered a two-sample case, but the approach considered here is more general and based on latent variable modeling. Although stratified procedures can be conceptualized in log-linear terms (Bishop, Fienberg, &

Holland, 1975), the present approach exploits a result from Dayton and Macready (1983) who showed that the model underlying the McNemar test is equivalent to a restricted two-class latent class model for a 2×2 contingency table.

Latent Class Analysis

The mathematical model for latent class analysis (LCA) can be conceptualized as follows. Let $Y_s = \{y_{sj}\}$ be the vector-valued response for observed variables $j = 1, \ldots, J$, for the s^{th} respondent. Let the response options for the variables be defined over a set of distinct, mutually-exclusive values $r = 1, \ldots, R_j$ for the j^{th} variable (e.g., for dichotomous responses these values would be $r = (1,2)$). Then, for C distinct latent classes, an unrestricted latent class model is defined as:

$$P(Y_s) = \sum_{c=1}^{C} \theta_c \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr}^{\delta_{sjr}} \ .$$

The latent class (mixing) proportions are θ_c, $c = 1, \ldots, C$, with the restriction that these non-negative proportions sum to one. The latent class proportions represent the sizes of the unobserved latent classes. The α_{cjr} are conditional probabilities associated with the observed variables. That is, they represent the probability of response r to variable j given membership in the c^{th} latent class. Thus, for each variable, there is a vector of R_j conditional probabilities and these conditional probabilities sum to one for each variable within each latent class.

The δ_{sjr} terms are introduced in the manner of Kronecker deltas to include the appropriate conditional probabilities in the model based on the observed responses for the s^{th} respondent. Thus, $\delta_{sjr} = 1$ if $y_{sj} = r$ but $\delta_{sjr} = 0$ otherwise. In effect, the latent class model is based on the assumption that, conditional on latent class membership, the responses to the variables are independent. To make the model explicit, consider three dichotomously-scored variables and two latent classes. Within latent class 1, the probabilities for a 1 response (e.g., positive, yes or agree) are α_{111}, α_{121}, and α_{131} and within latent class 2 these probabilities are α_{211}, α_{221}, and α_{231}. The observed response {1,2,1}, for example, has conditional probability $\alpha_{111}\,(1 - \alpha_{121})\,\alpha_{131}$ within latent class 1 and conditional probability $\alpha_{211}\,(1 - \alpha_{221})\,\alpha_{231}$ within latent class 2, so that the unconditional probability for this response is $\theta_1\alpha_{111}\,(1 - \alpha_{121})\,\alpha_{131} + (1 - \theta_1)\,\alpha_{211}\,(1 - \alpha_{221})\,\alpha_{231}$. From a psychological measurement perspective, each conditional probability can be viewed as an item difficulty (or easiness) that may vary across the unobserved latent classes.

The log-likelihood for a latent class model with observations, $Y_s = \{y_{sj}\}$, is $\Im = \sum_s LnP(Y_s)$. To generate maximum-likelihood estimates (MLEs) for the parameters in the model, a set of normal equations must be solved simultaneously: $\frac{d\Im}{d\theta_c} = 0$ for each latent class proportion and $\frac{d\Im}{d\alpha_{cjr}} = 0$ for each conditional probability. However, a specific model will involve restrictions that must be introduced into the solution for the estimates. For example, the latent class proportions must sum to 1 across the classes and the conditional probabilities may be constrained in various ways including, at least, summing to 1 across the response options. Unfortunately, the presence of additive terms within the logarithmic operator means that the model is non-linear in the parameters and, except for special cases, cannot be solved by algebraic approaches.

However, given suitable restrictions, maximum-likelihood estimation is usually possible using iterative procedures such as Newton-Raphson algorithms as in Haberman's program LAT (1979) or by estimation-maximization (EM) algorithms as in Vermunt's program LEM (1997). These procedures are *regula falsi* methods that are subject to various computing complications including local maxima, boundary conditions, etc. (Dayton, 1999). Based on the MLE's, model fit can be assessed by Pearson or likelihood-ratio chi-square statistics computed from the cross-tabulation of the observed responses (e.g., the 2^J table for J dichotomous variables). In general, the degrees of freedom for these tests are $\#Cells - 1 - \#Pars$ where $\#Pars$ is the number of independent parameters estimated by MLE. However, it is possible that the parameters in a latent class model are not identified even though there are positive degrees of freedom. Programs such as LEM (Vermunt, 1997) provide some useful information on model identification although this can be a complex issue. These methods, as well as related descriptive approaches to assessing model fit, are summarized in Dayton (1999).

Two Repeated Dichotomous Classifications

The McNemar test is based on a 2×2 table with observed cell frequencies n_{ij} and cell proportions $p_{ij} = n_{ij} / N$ where N is the total sample size. Assuming an unrestricted two-class latent class model, the expected cell proportions are:

$$E(p_{11}) = \theta_1\alpha_{111}\alpha_{121} + \theta_2\alpha_{211}\alpha_{221}$$
$$E(p_{12}) = \theta_1\alpha_{111}\alpha_{122} + \theta_2\alpha_{211}\alpha_{222}$$
$$E(p_{21}) = \theta_1\alpha_{112}\alpha_{121} + \theta_2\alpha_{212}\alpha_{221}$$
$$E(p_{22}) = \theta_1\alpha_{112}\alpha_{122} + \theta_2\alpha_{212}\alpha_{222}$$

Given the usual restrictions on probabilities, there are five independent parameters, θ_1, α_{111}, α_{121}, α_{211}, and α_{221}, but only three independent observed proportions, p_{11}, p_{12}, and p_{21}. Therefore, the model cannot be identified unless at least two more restrictions are imposed. Imposing two restrictions would not yield positive degrees of freedom for assessing fit, so, in order to assess fit of the model, a total of three additional restrictions is required. The first two restrictions can be: $\alpha_{111} = \alpha_{121} \equiv \alpha_{11}$ and $\alpha_{211} = \alpha_{221} \equiv \alpha_{21}$; i.e., equating conditional probabilities across the two variables. If we interpret the first class as favoring a "1" response and the second class as favoring a "2" response, then a third restriction of the form $1 - \alpha_{11} = \alpha_{21} \equiv \alpha_e$ allows a single conditional probability, α_e, to be viewed as a response error. It should be noted that Proctor (1970) suggested the use of a restricted latent class model that involved response errors for the analysis of Guttman scales and that his approach was expanded by Dayton and Macready (1976). Given these restrictions, the equations above reduce to:

$$E(p_{11}) = \theta_1(1-\alpha_e)^2 + (1-\theta_1)\alpha_e^2$$
$$E(p_{12}) = \theta_1(1-\alpha_e)\alpha_e + (1-\theta_1)\alpha_e(1-\alpha_e)$$
$$E(p_{21}) = \theta_1\alpha_e(1-\alpha_e) + (1-\theta_1)(1-\alpha_e)\alpha_e$$
$$E(p_{22}) = \theta_1\alpha_e^2 + (1-\theta_1)(1-\alpha_e)^2$$

The two latent classes can be interpreted as comprised of respondents who consistently use the response category 1 or, alternately, consistently use the response category 2. Inconsistent responses such as {1,2} or {2,1} are assumed to occur as a result of response errors that represent lack of consistency. Note responses such as {1,1} and {2,2} require that respondents either do not make a response error or that they make two response errors (e.g., a respondent in the latent class associated with a {1,1} response makes two response errors and responds {2,2}).

For this relatively simple model, the log-likelihood and normal equations can be set up and solved algebraically as shown in Dayton and Macready (1983).

However, an alternative approach is based on the realization that the expected and observed frequencies are equal for responses {1,1} and {2,2}; i.e., $p_{11} = E(p_{11}) = \theta_1(1-\alpha_e)^2 + (1-\theta_1)\alpha_e^2$ and $p_{22} = E(p_{22}) = \theta_1\alpha_e^2 + (1-\theta_1)(1-\alpha_e)^2$. Thus, algebraically solving these two equations for values of the parameters yields, per force, the maximum likelihood estimators:

$$\hat{\theta}_1 = \frac{p_{11} - \hat{\alpha}_e^2}{1 - 2\hat{\alpha}_e} \text{ and } \hat{\alpha}_e = .5 - \sqrt{.25 - (p_{12} + p_{21})/2}\,.$$

Note that $\hat{\alpha}_e$ is undefined for $p_{12} + p_{21} > .5$ so that it is necessary to reverse the coding for one of the variables if this occurs in practice. The restricted latent class model yields expected frequencies that are consistent with the McNemar test in the sense that $\hat{p}_{11} = p_{11}$, $\hat{p}_{22} = p_{22}$, and $\hat{p}_{12} = p_{21} = (p_{12} + p_{21})/2$. Also, the resulting chi-square value for model fit is exactly the same as the uncorrected McNemar chi-square statistic with one degree of freedom. Thus, the McNemar may be viewed as testing the null the hypothesis $\alpha_{11} = 1 - \alpha_{21}$ versus the alternative $\alpha_{11} \neq 1 - \alpha_{21}$.

This conceptualization of the McNemar test focuses on response consistency rather than marginal homogeneity although the implications for observed responses are the same. However, estimates for the latent class parameters provide a measure of the agreement between classifications that is not available in a conventional McNemar analysis. For example, consider the exemplary before/after treatment results in Table 2. Positive responses occur at a rate of 40.3% before treatment and at a rate of 47.6% after treatment. The 6.3% difference is significant based on an uncorrected McNemar chi-square value of 4.55 (p = .033). Our latent class model yields estimated parametric values of .423 for the latent class proportion, θ_1, and .074 for the error rate, α_e. The value .423, or 42.3%, is an estimate for the proportion of respondents who have positive responses at both the before and after occasions of observation. Note that the conventional McNamar procedure does not provide a comparable statistic. Also, the value .074, or 7.4%, is an estimated error rate that applies to both the positive/positive and negative/negative latent response groups. Once again, this a value that has no direct analog in a McNemar analysis (although roughly similar to the before/after relative change in this example).

Table 2. Exemplary Pre/Post Data

		After		
		Positive	***Negative***	***Total***
Before	***Positive***	59	6	65
	Negative	16	80	96
	Total	75	86	161

Stratified McNemar Test

Consider cross-tabulations similar to those in Table 1 for two or more strata within a population (or for the same population at different points in time or for samples from several populations). Letting the strata be represented by $y = 1, \ldots, Y$, the expected cell proportions for a given stratum can be written as:

$$
\begin{aligned}
E\left(p_{11y}\right) &= \theta_{1y}\left(1-\alpha_{ey}\right)^2 + \left(1-\theta_{1y}\right)\alpha_{ey}^2 \\
E\left(p_{12y}\right) &= \theta_{1y}\left(1-\alpha_{ey}\right)\alpha_{ey} + \left(1-\theta_{1y}\right)\alpha_{ey}\left(1-\alpha_{ey}\right) \\
E\left(p_{21y}\right) &= \theta_{1y}\alpha_{ey}\left(1-\alpha_{ey}\right) + \left(1-\theta_{1y}\right)\left(1-\alpha_{ey}\right)\alpha_{ey} \\
E\left(p_{22y}\right) &= \theta_{1y}\alpha_{ey}^2 + \left(1-\theta_{1y}\right)\left(1-\alpha_{ey}\right)^2
\end{aligned}
$$

Maximum likelihood estimation for the stratified model follows the same approach as for any latent class model in general but requires that suitable restrictions be imposed on the estimated parameters. In addition, issues related to identification of the model must be considered (Dayton, 1999). Because the strata are independent, it is apparent that jointly estimating the parameters in the heterogeneous form of the stratified model is the same as fitting the model separately to each stratum but does provide an overall measure of fit in the form of a chi-square statistic with Y degrees of freedom. However, the major advantage of conceptualizing the model in this form is that it allows for imposing across-strata restrictions on the error rates. The most highly restricted case results in a homogeneous model with $2Y - 1$ degrees of freedom that is based on restrictions of the form $\alpha_{ey} = \alpha_e \forall y$. However, a variety of part-heterogeneous models may be suggested by theory (or, the data) and tested accordingly. Closed-form estimates are not, in general, available for the stratified model. Fortunately, as illustrated below, available programs for latent class analysis allow for these restrictions and associated MLEs.

A similar conceptualization, known as the Hui-Walter model (Hui & Walter, 1980), has been presented in the context of repeated assays for the purpose of estimating false-positive and false-negative rates. This model is saturated so that fit to data cannot be assessed by ordinary procedures and is based on a different set of restrictions. Biemer (2011) presents an extended discussion with examples of the Hui-Walter model.

Application for Two Immunization Survey Items

The CDC Behavioral Risk Factor Surveillance System (BRFSS) is a large-scale telephone survey that tracks health risks in the United States. The CDC web-enabled analysis tool for BRFSS (http://nccd.cdc.gov/s_broker/WEATSQL.exe/weat/index.hsql) was used to produce cross-tabulations of responses to two items, referred to as Flu and Pneumonia, for adults aged 65 and older:

Flu:	Had a flu shoot within past 12 months.
Pneumonia:	Ever had a pneumonia vaccination.

The item responses were Yes/No and, for the year 2011, there were responses available for a total of 143,002 people across the United States. A large variety of demographic variables is included in the data system and, using CDC labeling, we chose to compare race/ethnicity groups divided into the strata: (1) White, Non-Hispanic; (2) Black, Non-Hispanic; (3) Hispanic; and (4) Other which comprised multiracial and other races. Cross-tabulated frequency data for the four race/ethnicity groups are presented in Table 3.

Table 3. Cross-Tabulation of Two Immunization Variables for Four Race/Ethnic Groups

Flu:	**Yes**	**Yes**	**No**	**No**		**McNemar**	
Pneumonia:	**Yes**	**No**	**Yes**	**No**	**Total**	**G^2**	**Prob.**
White, Non-Hispanic	64,446	12,729	23,792	21,279	122,246	3404.05	0.000
Black, Non-Hispanic	3,367	1,107	1,728	2,575	8,777	137.14	0.000
Hispanic	2,050	1,005	1,123	2,251	6,429	6.55	0.011
Other	2,641	679	1,105	1,125	5,550	102.71	0.000
Total	*72,504*	*15,520*	*27,748*	*27,230*	*143,002*	*3503.30*	*0.000*

Our focus was on the relative rates of flu and pneumonia immunizations across the race/ethnic groups. As shown in Table 4, the marginal immunization

rates are moderately different for three of the four race/ethnic groups but very similar for Hispanics (i.e., .48 and .49 for flu and pneumonia, respectively).

Table 4. Marginal Rates

Race/Ethnic Group	Flu	Pneumonia
White, Non-Hispanic	0.63	0.72
Black, Non-Hispanic	0.51	0.58
Hispanic	0.48	0.49
Other	0.60	0.67
Total	*0.62*	*0.70*

In Table 3, the column labeled McNemar G^2 presents McNemar likelihood-ratio chi-square fit statistics for each race/ethnic group as well as for the total sample. These tests are consistent with our observation concerning the marginal rates with only the Hispanic group failing to be significant beyond the .01 level.

Homogeneous, heterogeneous and part-heterogeneous stratified McNemar models were fit to the cross-tabulations of the two immunization items for the four race/ethnic groups. The homogeneous model posits a single response error rate, α_e, for the four strata whereas the heterogeneous model posits unique error rates, α_{e1}, α_{e2}, α_{e3}, and α_{e4}, for the four strata. In both cases, the size of the latent class, θ_1, corresponding to a Yes response to both items, {1, 1}, is allowed to vary by group in order to fix the marginal distributions for the race/ethnic groups. The part-heterogeneous model, which equated error rates for all groups except White, Non-Hispanic, was suggested by the fact that the error rates for these three strata were quite similar for the heterogeneous model (i.e., .206, .209 and .201, respectively). MLE parameter estimation and model fit were conducted using the latent variable program, LEM (Vermunt, 1997). Although lacking a modern computer interface, LEM has the dual advantages of being (a) available free for download for Microsoft operating systems and (b) extremely flexible in terms of the latent class models that can be estimated. Sample LEM program set-ups for the homogeneous and heterogeneous models are included in the Appendix. Model fit statistics and parameter estimates are presented in Table 5. Given the large sample size, it was not unexpected that all three models result in rejection of the hypothesis of equal error rates across the four race/ethnic groups.

Table 5. Stratified McNemar Models Fit to Vaccination Variables

Model	DF	Chi-Sq (G^2)	AIC	Homogenous Groups	Error Rates	Class Size
Homogeneous	7	3709.95*	513, 360.7	[1234]	.186	.78, .57, .47, .72
Part-Heterogeneous	6	3652.38*	513, 305.1	[1],[234]	.183, .204	.78, .58, .47, .73
Heterogeneous	4	3650.85*	513, 307.6	[1],[2],[3],[4]	.183, .206,.209, .201	.77, .58, .47, .73
Collapsed	1	3503.30*	N/A	[1]	.186	.75

Note: *All *p*-values are less than .001

Using the Akaike (1973) information measure as suggested by Dayton (1999) for comparing latent class models, a min(AIC) criterion indicates that the part-heterogeneous model is best among the models being compared. Because the three models are nested, it is appropriate to test differences among them using likelihood-ratio chi-square (G^2) statistics. These comparisons are:

Homogeneous vs. Part-Heterogeneous: $\Delta(G^2) = 57.57$, DF = 1, $p < .01$;
Homogeneous vs. Heterogeneous: $\Delta(G^2) = 59.10$, DF = 3, $p < .01$;
Part-Heterogeneous v.s Heterogeneous: $\Delta(G^2) = 1.53$, DF = 2, $p < .05$.

The Part-Heterogeneous model fits the data no worse than the Heterogeneous model, whereas both of these models provide better fit than the Homogeneous model.

As noted above, in order to fix the marginal distributions at observed values for the four race/ethic groups, it was necessary to posit separate latent class proportions for the strata. These proportions are quite consistent across the models that were evaluated with White, Non-Hispanic and Hispanic showing considerably larger latent class proportions than the other two groups. If race/ethnicity is ignored and a non-stratified latent class model is fitted to the (marginal) 2×2 table of immunization rates, a latent class proportion of .75 is estimated. An error rate of .186 was estimated for the homogeneous model which is essentially identical to that from the marginal 2×2 model although this is driven by the fact that about 85% of the total sample is comprised of White, Non-Hispanic respondents,

In order to allow for the observed lack of agreement in immunizations rates for flu and pneumonia vaccinations, the latent class models suggest a rate of inconsistencies (errors) of approximately 18% - 20%. That is, about one in five individuals in a latent class that represents consistently Yes (or consistently No) respondents would, in fact, fail to respond consistently. From Table 2 it is notable

that inconsistencies tend to be in the direction of failing to obtain a flu vaccination, which may suggest some educational strategy in this regard for the 65 and older age group.

Capitalizing on the fact that the McNemar test can be conceptualized as a restricted latent class model, we have defined homogeneous, heterogeneous and part-heterogeneous models with parameter estimates that have interpretations that could be of interest in applied research settings such as immunization patterns for the 65-and-over population. Furthermore, estimation and significance testing are available using widely available latent-class programs.

References

Agresti, A., & Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Applied Statistics, 54*(4), 691-706. doi: 10.1111/j.1467-9876.2005.05437.x

Akaike, H. (1973). Information theory and an extension of the maximum-likelihood principle. In B. N. Petrov and F. Csake (Eds.), *Second international symposium on information theory*. Akademiai Kiado: Budapest, 267-281.

Biemer, P. P. (2011). *Latent class analysis of survey error*. New Jersey: Wiley.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Dayton, C. M. (1999). *Latent class scaling analysis*. New York: Sage Publications. doi: 10.4135/9781412984720

Dayton, C. M. & Macready, G. B. (1976). A probabilistic model for the validation of behavioral hierarchies. *Psychometrika, 41*(2), 189-204. doi: 10.1007/bf02291838

Dayton, C. M. & Macready, G. B. (1983). Latent structure analysis of repeated classifications with dichotomous data. *British Journal of Mathematical & Statistical Psychology, 36*(2), 189-201. doi: 10.1111/j.2044-8317.1983.tb01124.x

Durkalski, V. L., Palesch, Y. Y., Lipsitz, S. R. & Rust, P.F. (2003). Analysis of clustered matched-pair data. *Statistics in Medicine, 22*(15), 2417-2428. doi: 10.1002/sim.1438

Fagerland, M. W., Lydersen, S. & Laake, P. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology, 13*(1), 91. doi: 10.1186/1471-2288-13-91

Feuer, E. J. & Kessler, L. J. (1989). Test statistic and sample size for a two-sample McNemar test. *Biometrics, 45*(2), 629–636. doi: 10.2307/2531505

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Haberman, S. J. (1979). *Analysis of qualitative data, volume 2: New developments*. New York: Academic Press.

Hui, S. L. & Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics, 36*(1), 167-171. doi: 10.2307/2530508

Klingenberg, B. & Agresti, A. (2006). Multivariate extensions of McNemar's test. *Biometrics, 62*(3), 921-928. doi: 10.1111/j.1541-0420.2006.00525.x

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry, 116*(535), 651-655. doi: 10.1192/bjp.116.535.651

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157. doi: 10.1007/bf02295996

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika, 35*(1), 73-78. doi: 10.1007/bf02290594

Stuart, A. A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42*(3-4), 412-416. doi: 10.1093/biomet/42.3-4.412

Vermunt, J. K. (1997). *LEM: A general program for the analysis of categorical data*. Department of Methodology & Statistics, Tilburg University.

Westfall, P. H., Troendle, J. F. & Pennello, G. (2010). Multiple McNemar Tests. *Biometrics, 66*(4), 1185-1191. doi: 10.1111/j.1541-0420.2010.01408.x

Appendix

LEM input file for Homogeneous model

```
* CDC Behavioral Risk Factor Surveillance System
* Elderly flu shot last 12 months
* Elderly pneumonia vaccination ever
* Four ethnic groups - white, black, Hispanic, other
* Stratified McNemar test
* Homogenous Model [1234]
lat 1
man 3
dim 2 4 2 2
lab X Y F P * X = latent variable; Y = Ethnic;
              F = Flu, P = Pneumonia
mod Y
X|Y
F|XY eq2
P|XY eq2
des [ 0 2 0 2 0 2 0 2  2 0 2 0 2 0 2 0
      0 2 0 2 0 2 0 2   2 0 2 0 2 0 2 0 ]
dat [64446 12729 23792 21279     3367 1107 1728 2575
      2050  1005  1123  2251     2641  679 1105 1125]
```

LEM input file for Heterogeneous model

```
* CDC Behavioral Risk Factor Surveillance System
* Elderly flu shot last 12 months
* Elderly pneumonia vaccination ever
* Four ethnic groups - white, black, Hispanic, other
* Stratified McNemar test
* Heterogeneous Model [1],[2],[3],[4]
lat 1
man 3
dim 2 4 2 2
lab X Y F P * X = latent variable; Y = Ethnic;
              F = Flu, P = Pneumonia
mod Y
      X|Y
```

```
       F|XY eq2
       P|XY eq2
des [ 0 2 0 4 0 6 0 8   2 0 4 0 6 0 8 0
      0 2 0 4 0 6 0 8   2 0 4 0 6 0 8 0 ]
dat [64446 12729 23792 21279     3367 1107 1728 2575
      2050  1005  1123  2251     2641  679 1105 1125]
```

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 34-50.
doi: 10.22237/jmasm/1493596920

ISSN 1538 − 9472

An Empirical Demonstration of the Need for Exact Tests

Vance W. Berger
National Cancer Institute
Rockville, MD

The robustness of parametric analyses is rarely questioned or qualified. Robustness, generally understood, means the exact and approximate p-values will lie on the same side of alpha for any reasonable data set; and 1) any data set would qualify as reasonable and 2) robustness holds universally, for all alpha levels and approximations. For this to be true, the approximation would need to be perfect all of the time. Any discrepancy between the approximation and the exact *p*-value, for any combination of alpha level and data set, would constitute a violation. Clearly, this is not true, and when confronted with this reality, the "No True Scotsman" fallacy is often invoked with the declaration it must have been a pathological data set, as if this would obviate the responsibility to select an appropriate research method. Ideally, a method would be selected because it is optimal, or at least appropriate, without needing special pleading, but judging by how often approximations are used when the exact values they are trying to approximate are readily available, current trends do not come close to this ideal. One possible explanation might be that there is not much information available on data sets for which the approximations fail miserably. Examples are presented in an effort to clarify the need for exact analyses.

Keywords: Chi-square test, normality, permutation tests, robustness, t-test

Introduction

Approximations are used rather often, in all sorts of contexts. Sometimes this is because the exact value is not available, or because it could be made available but only at a prohibitive cost. In no case is the approximation ever actually preferred to the exact value it is trying to approximate, for if this indeed is the case, then the approximation is not an approximation. Rather, it would then be calculated for its inherent interest.

This raises the issue of whether parametric analyses are conducted because they are of interest in their own right, or merely as approximations to exact

Vance W. Berger, PhD, is a member of the Biometry Research Group. Email him at: vb78c@nih.gov.

analyses. Though it is conceivable that in certain limited cases there is interest in a parametric analysis, it is clear, when one considers the pre-testing that generally occurs to ensure that the conditions are met to ensure the integrity of the approximation, that the parametric analyses are, in general, just approximations, nothing more. For example, if one were to test the data for normality by any method, even an informal one such as appeal to the fact that we have always just assumed normality, prior to conducting a *t*-test, then this undermines the notion that the *t*-test is conducted for inherent interest. There is interest only conditionally on the finding that the data are normal enough to merit such interest. Along these lines, Bradley (1968) noted that "A corresponding parametric test is valid only to the extent that it results in the same statistical decision [as the exact test]" (p. 85).

We must distinguish two cases here. In one case, the choice is to approximate or not to approximate; but if one does, then one cannot know how well the approximation performed since the exact value cannot be computed. In the other case, the exact value is readily available, so here the choice is to use it or the approximation. Berger (2000) pointed out the folly, in this case, of ever using the approximation. After all, how compelling is a test of normality in allowing for the use of an approximation when one can instead simply compare the two values to see how close they actually are (as opposed to how close they *should* tend to be on average)? But for that matter, given that one already has the exact value, why even consider replacing it with the approximation?

The lapse in logic that would allow a researcher to use an approximation when the very quantity it is trying to approximate is readily available is staggering, and yet this exact situation plays out in a huge number of randomized clinical trials, Bradley's aforementioned sage wisdom notwithstanding. The randomization itself allows for exact comparisons of the treatment groups by way of permutation tests (see, e.g., Fisher, 1935; Rigdon & Hudgens, 2015; Lu, Ding, & Dasgupta, 2015), and yet it is the inexact parametric tests that are used far more often, generally after going through the motions of justifying this choice by first conducting a test of the assumptions that allegedly support the use of the parametric test in question.

The only saving grace would be if it just didn't matter. Sure, the exact analyses are preferable, but given how robust the parametric analyses are, there is very little to gain and much to lose in terms of computing time. This argument may have been compelling decades ago, when it actually would have been difficult to conduct a permutation test, but today this is no longer the case. It is just as easy to do it right as it is to do it wrong. So this leaves us at the other

aspect of this argument, it just doesn't matter (and all the variations of this theme, including the assertion that there are more important issues for statisticians to concern themselves with, as if the choice of an appropriate analysis is somehow beneath the dignity of the very party charged with doing so). Moreover, even if it did not matter (at least numerically), that still would not provide a compelling argument in favor of a theoretically unsound analysis.

This much is clear, and should already suffice to eradicate parametric analyses from actual clinical trials, at least when comparing treatments. Sadly, it has not, and the widespread delinquency of researchers who simply cannot be bothered to concern themselves with the relative merits of various analyses is matched by a commensurate delinquency on the part of those authorities who could impose the need for rigor, yet somehow choose not to. And they do this while assuring patients and funding bodies that only the best research methods will be used. But at least we can fall back on robustness.

Everybody knows that parametric analyses are robust, but how many can actually provide a precise formulation of what that means, operationally? How good is good enough? What does "good enough" even mean in this case? What does convergence as the sample size increases without bound say about the discrepancy for this particular data set with its very finite sample size? These are uncomfortable questions for those who continue to embrace robustness as a justification for using approximations when in fact the exact values should be used instead. One theorem that would be useful in supporting this case would be along the lines of $|p_1 - p_2| < k/n$, where k is some universal constant, n is the sample size, and k/n bounds the absolute difference between the two p-values.

Even if this statement were true, it would still be hard to see how that would justify the substitution of the one for the other. After all, enlightened researchers recognize that each party may apply his or her own personal alpha level to the results of any clinical trial (Berger, 2004). This being the case, how much error is acceptable when, with a different choice, we can attain the ideal of no error at all? Moreover, is such a bound of the discrepancy even true? The remainder of this paper will illustrate that in fact it is not true for any reasonable value of k. We will consider the chi-square approximation to Fisher's exact test, the Smirnov test (both exact and approximate), and the t-test in the sections to follow.

Examples of the Chi-Square Test Failing

When dealing with a single 2×2 contingency table, the two most common tests seem to be Fisher's exact test and the chi-square test. Of course, the chi-square

test is used in other situations as well, and sometimes the exact test to which it is compared is not Fisher's exact test, and in some cases this test may not even have a name (but is easily defined in terms of a test statistic and a permutation mode of inference). Table 1 presents six data sets for which the chi-square *p*-value differs markedly from its exact counterpart. In Example C1, the comparison was the chi-square test to Fisher's exact test. Little (1989) pointed out that each expected cell count was over five, so the usual rule of thumb would have led one to use the chi-square test and find significance at the 0.05 level (note that the *p*-values in the table are one-sided, so Fisher's exact test is not significant).

Table 1. Data sets for which the chi-square test fails badly

N	References	Data Set*	*p*-values**
C1.	Little (1989)	{(170,2);(162,9)}	0.0299, 0.0162
C2.	Zelterman et al. (1995)		0.0424, 0.119
C3.	Cytel Software (1995, p. 11)		0.0013, 0.1342
C4.	Cytel Software (1995, p. 17)	{(3,1);(1,3)}	0.243, 0.0786
C5.	Berger and Lachenbruch (1998)	{(20,230);(35,225)}	0.063, 0.047
C6.	Hewett et al. (1999); Clancy (2000)	{(10,453);(2,364)}	NS***, 0.02

Note: Citations abbreviated for space; see Reference section below for full reference
* Data set provided only for a single 2 × 2 contingency table
** Exact *p*-value first, then chi-square *p*-value
*** Actual *p*-value not reported, nor is the full data set available

Table 2. Data from StatXact (Cytel Software, 1995)

0	7	0	0	0	0	0	1	1
1	1	1	1	1	1	1	0	0
0	8	0	0	0	0	0	0	0

Example C2 is from Table 1 of Zelterman, Chan, and Mielke (1995), which is hypothetical data in the form of two stratified 2 × 2 contingency tables. These were $\{(1, 0); (3, 9)\}$ and $\{(0, 0); (9, 5)\}$. Not only do the *p*-values differ dramatically (the exact *p*-value is 0.0424 and the approximate chi-square *p*-value is 0.119), but in fact it is the exact one that is lower. This example flies in the face of the conventional wisdom that states that permutation tests are always conservative so therefore exact *p*-values are always larger than their approximate counterparts. Zelterman et al. (1995) note "The lesson we learn ... is that the behavior of test statistics, such as Pearson's chi-square, may or may not agree with their asymptotic approximations. The only certain methods for accurate analysis of tables with small counts is to perform exact methods based on the

likelihood function" (p. 358) Example C3 is based on a sparse 3 × 9 contingency table presented in the StatXact manual (Cytel Software, 1995, p. 11), which is reproduced in Table 2.

Pearson's chi-square test of an interaction between rows and columns has a test statistic value of 22.29 with (3 – 1)(9 – 1) = 16 degrees of freedom, for a *p*-value of 0.1342. Using the same test statistic, specifically the chi-square test statistic, but using its exact distribution instead of the distributional assumption results in an exact *p*-value of 0.0013. As in Example C2, not only are the *p*-values (and the interpretations one would arrive at) grossly different from each other, but in fact it is the exact one that would demonstrate a true treatment effect (assuming that rows are treatments), whereas the approximate one would miss it. The StatXact manual notes "the need to compute the exact *p*-value, rather than relying on asymptotic results, whenever the data set is small, sparse, unbalanced, or heavily tied. The trouble is that it is difficult to identify, a priori, that a given data set suffers from these obstacles to asymptotic inference" (Cytel Software, 1995, p. 11).

Example C4 is also from the StatXact manual (Cytel Software, 1995, p. 17), and is Fisher's famous original tea-tasting experiment which led to the development of Fisher's exact test. As is well known, the experiment involved testing the claim of a British woman that she was able to distinguish between the two possible orders, milk first and then tea, or tea first and then milk, being poured into a cup. This woman was presented with eight cups of tea, in which four were of each order (and she was told this key fact). The order in which the cups were given to her was randomized. Of the four cups with milk poured first, she guessed right three times. Likewise, of the four cups with tea poured first, she guessed right three times. The chi-square test yields a *p*-value of 0.1573 two-sided or 0.0786 one-sided. The Fisher exact *p*-value is 0.243, which is not even close.

Example C5 regards data presented at the December 15, 1995 FDA Blood Products Advisory Committee meeting. Hospitalization due to a targeted respiratory disease was required by 20/250 (8.0%) patients on a biological treatment arm and 35/260 (13.5%) patients on the control arm. Pearson's uncorrected chi-square test yielded $p = 0.047$ two-sided, and significance was declared at the prospectively specified 0.05 alpha level (two-sided). But the nominal 0.05 alpha level is preserved only if the true probability of a Type-I error is no greater than 0.05. A fair question, then, is how likely one would be to obtain data at least as significant ($p < 0.047$), by using this chi-square test, assuming nothing more than random allocation of patients to treatment groups. The answer, $p = 0.063$, is provided by Fisher's exact test, which of course does not attain

statistical significance at the 0.05 alpha level. The StatXact manual points out that "The term 'asymptotically' means 'given a sufficient sample size', though it is not easy to describe the sample size needed for the chi-square distribution to approximate well the exact distribution of the Pearson statistic" (Cytel Software, 1995, p. 12).

Example C6 is based on Clancy's (2000) letter to the editor regarding Hewett, Lindenfeld, Riccobene, and Noyes' (1999) paper, in which the authors evaluated the effect of neuromuscular training on the incidence of knee injury in female athletes. There were ten injuries among 463 untrained athletes and two injuries among 366 trained athletes. The chi-square test was reported to yield $p = 0.02$. Clancy reported a non-significant p-value with Fisher's exact test, and also pointed out that one cell had both an actual and an expected cell count under five, so that Fisher's exact test would be the more reliable of the two, in keeping with conventional wisdom. Notably, Hewett, Levy, and Noyes (2000) responded to the letter by resorting to appeal to credentials, stating essentially that they used an "excellent" statistician, so therefore whatever he came up with must be correct by virtue of his coming up with it. A second "unbiased" statistician confirmed this.

Even in the absence of a reason for suspicion, suspicion must still arise when an argument is defended by appeal to credentials. This is, after all, tantamount to an admission that there is no better defense for the argument than credentials. One has to wonder just how "unbiased" the second statistician truly was, and also how many competent statisticians (with the fortitude to refuse to sign off on an analysis so poorly planned) were also contacted. Competent statisticians know to use Fisher's exact test when the expected cell counts, or any one of them, is less than five; even better statisticians would recognize the irrelevance of the expected cell counts and instead use Fisher's exact test any time it differs substantially from the chi-square test. And still better statisticians would recognize that they are not in a position to determine how close an approximation needs to be in order that it be preferred to the quantity it is trying to approximate, so they would simply use Fisher's exact test routinely.

Examples of the Approximate Smirnov Test Failing

When dealing with a single ordered $2 \times J$ table, the best test that is offered as a routine option (no programming required) in commercially available software packages is the exact Smirnov test, a standard feature of StatXact. See Section 10.1 of Hollander and Wolfe (1973) and Section 1.6 of Lehmann (1975). Note that while it is customary to speak of the Smirnov test as a two-sided approximate

test, we use this term to denote the exact one-sided version. Essentially, the only difference between the one-sided and the two-sided version is the absence or presence, respectively, of absolute values around the directed difference of CDFs to be maximized. Whether one-sided or two-sided, the approximate test that bears the same name often gives strikingly different p-values from the exact version for the same data set, as we will demonstrate in Table 3. Note that the exact Smirnov test *p*-values (but not the approximate ones) for these data sets appeared in Table 2 of Berger (2002), and some of them seem to contradict what we are presenting now in our Table 3. The reason for this is the newfound ability of StatXact to compute exact Smirnov *p*-values immediately for such large data sets, whereas only a few years ago only Monte Carlo approximations were feasible.

Table 3. Ordered 2 × *J* tables for which the approximate Smirnov test fails badly

N	References	Data Set	*p*-values*
S1.	Fentiman et al. (1983)	{(6,8,4,2,3);(3,2,8,0,10)}	0.0138, 0.0296
S2.	Fox et al. (1993)	{(1,5,16);(0,0,22)}	0.0106, 0.1947
S3.	Fox et al. (1993)	{(12,3,7);(3,7,12)}	0.0108, 0.0252
S4.	Elwood (1998)	{(33,5,545);(29,8,836)}	0.0258, 0.6823
S5.	TOAST (1998)	{(291,168,176);(270,161,215)}	0.0379, 0.1376
S6.	Clark et al. (1999)	{(207,19,80);(181,25,101)}	0.0209, 0.0988
S7.	Clark et al. (1999)	{(187,15,104);(169,32,106)}	0.0938, 0.3242
S8.	Shelton et al. (2001)	{(83,14,5);(72,12,14)}	0.0766, 0.4147
S9.	Staszewski et al. (2001)	{(149,29,104);(144,15,121)}	0.1051, 0.3238

Note: Citations abbreviated for space; see Reference section below for full reference
* Exact one-sided Smirnov *p*-value first, then the approximate one-sided Smirnov *p*-value

Notice that in each case the approximate *p*-value is much larger than its exact counterpart. This refutes the common misunderstanding that exact *p*-values are always overly-conservative and therefore larger than the approximate *p*-values they would (and should) replace. Example S1 comes from a study of talc for malignant pleural effusions. There were 46 patients, and 23 were randomized to each group: talc and mustine. Some patients were considered to be "not assessable" because they died within a month of pleurodesis. Among the other patients (who were assessed), success or failure was defined in terms of radiologic criteria of effusion control. In addition to this binary success endpoint, patients were also classified as being alive or dead at the time the article was written, and as having had or not had evidence of recurrent effusion. So all in all we have four binary endpoints:

1. Died prior to assessment or not;
2. Dead or alive at the end of the study;
3. Success or not;
4. Recurrence or not.

This would appear to give $2 \times 2 \times 2 \times 2 = 16$ outcomes, but in fact the first two binary endpoints are fusible, because being alive at the end of the study necessarily entails also being alive long enough to be assessed. So instead of $2 \times 2 = 4$ outcomes for the first two binary endpoints above, we recognize the structural zero (one cannot die prior to being assessed and also be alive at the end of the study), and remove it to create a trichotomous information preserving composite endpoint, or IPCE, (died prior to assessment, assessed but dead at study end, alive at study end). See Berger (2002) for more information on the construction of the IPCE. We also note that the two binary endpoints success (yes/no) and recurrence (yes/no) are fusible, because recurrence is possible only if success was achieved in the first place, so we again have a structural zero (one cannot recur without having succeeded in the first place). Removing it gives the IPCE (no success, success then recurrence, success without recurrence). We have gone from $2 \times 2 \times 2 \times 2 = 16$ possible outcomes to only $3 \times 3 = 9$. But in fact further savings is possible too, as becomes evident from inspection of Table 4.

Dying before assessment precludes the possibility of a success, so the two lower left cells, labeled "SZ" in Table 4, are structural zeros. We make the simplifying assumption that death supersedes recurrence, and so we equate the two cells labeled "3" in Table 4. The upper right cell labeled "RZ" was a random zero; that is, there could have been patients surviving without success, but as it turned out, none did. This leaves only five active outcomes, labeled 1-5 in Table 4:

1. Died prior to being assessed;
2. Died after being assessed but without success;
3. Died after success;
4. Alive at study end but recurred;
5. Alive at study end without recurrence.

These outcomes are, of course, in order of increasing clinical benefit, and the data, as presented in Table 3, were (6, 8, 4, 2, 3) in the mustine group and (3, 2, 8, 0, 10) in the talc group, and the one-sided (to show a benefit of talc in shifting to more favorable outcomes) Smirnov p-values were 0.0138 (exact) and

0.02955 (approximate). If one were to use the two-sided 0.05 alpha level and then cut it in half for a 0.025 one-sided alpha level (which seems to be lacking in any real basis, yet is still used quite often as a policy), then only the exact Smirnov test would show a statistically significant improvement in outcomes associated with talc.

Table 4. The construction of the IPCE for example S1

	Died Before Assessment	Assessed, then Died	Alive at Study End
No Success	1	2	RZ
Success, then Recurrence	SZ	3	4
Success, no Recurrence	SZ	3	5

Examples S2 and S3 both represent the same patients, with the same endpoint, with the same treatments. All that varies is the timing of the measurement. Specifically, Example S2 is Day 2 and Example S3 is Days 1-5, and both come from a study of combination therapy for nausea (Fox, Einhorn, Cox, Powell, & Abdy, 1993). What is so amazing is the complete reversal in the direction of the shift. The endpoint we consider is response, which is scored as complete, major, or none. Note that this endpoint is the IPCE of two component binary response endpoints presented by Fox et al., specifically the response rate and the complete response rate. Clearly, the two endpoints are fusible, since a complete response implies also a response.

At Day 2, the data were (1, 5, 16) in the ondansetron group and (0, 0, 22) in the combination (ondansetron plus dexamethasone plus chlorpromazine) group. In other words, there was absolutely no effect of the combination therapy for the response rate (22/22 vs. 21/22), but a fairly strong effect on the complete response rate (22/22 vs. 16/22). At the Days 1-5 assessment, the situation was reversed, with (12, 3, 7) in the ondansetron group and (3, 7, 12) in the combination group. Now there was not much of an effect of the combination therapy for the complete response rate (12/22 vs. 7/22), but a fairly strong effect on the overall response rate (19/22 vs. 10/22). Either binary endpoint would show significance at the 5% alpha level at one time point but not at the other, with one-sided Fisher's exact test *p*-values of 0.5000 for the Day 2 overall response rate, 0.0106 for the Day 2 complete response rate, 0.0049 for the Days 1-5 overall response rate, and 0.1116 for the Days 1-5 complete response rate. The exact Smirnov test yields one-sided *p*-values of 0.0106 (Day 2) and 0.0108 (Days 1-5). The approximate test yields

one-sided *p*-values of 0.1947 and 0.02518. Once again, only the exact Smirnov test shows statistical significance at the customary 0.025 one-sided level of significance.

Example S4 is reinfarction data, in which the reinfarction could be confirmed or not, or there could be no reinfarction at all. Each patient can be scored on an ordered categorical scale with three categories, (confirmed reinfarction, reinfarction not confirmed, no reinfarction). Note that once again this is the IPCE for two binary endpoints originally presented. The data for the two treatment groups (placebo, then sotalol) are presented in Table 3, and the Smirnov test was used to compare the groups. As can be seen, the asymptotic version of the test was way off, to the point of being almost unbelievable, relative to the exact Smirnov test. The exact and approximate one-sided *p*-values were 0.0258 and 0.6823. Note that a one-sided *p*-value is not, in general, half the corresponding two-sided *p*-value, and also that a one-sided *p*-value can exceed 0.5 if the trend is in the "wrong" direction. Of course, that is not the case with the data at hand, as we tested for the direction of sotalol being superior, and the data do trend in this direction. So it is unclear why the asymptotic test would behave this way. One must ask if the data themselves might suggest the need for the exact version of the test. Berger (2000) reports that "It is unclear how one would determine the advisability of the approximate test, but if one were to 'think unconditionally' then the small middle margin would not be a concern. The large sample sizes (over 500 per group), coupled with expected cell counts that all exceed five, would certainly be reassuring" (p. 1322).

Example S5 concerns danaparoid for acute ischemic stroke. The TOAST Investigators (The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment Investigators [TOAST], 1998) presented two binary endpoints, favorable outcomes (yes or no) and very favorable outcomes (yes or no), but again these two binary endpoints are clearly fusible, since a very favorable outcome implies also a favorable outcome. The IPCE is an ordered categorical outcome variable with categories for (unfavorable, favorable, very favorable). The TOAST Investigators inexplicably and indefensibly excluded some randomized patients from the analysis they called "intent-to-treat", but of course the correct intent-to-treat analysis would include all patients randomized. For now, we note that this set of patients can be classified by favorable outcomes at Day 7 as (291, 168, 176) in the placebo arm and (270, 161, 215) in the danaparoid group. The one-sided Smirnov *p*-values are 0.0379 (exact) and 0.1376 (approximate).

Examples S6 and S7 both come from the study of Clark et al. (1999) comparing rt-PA to placebo for ischemic stroke. The primary endpoint was a

complete recovery, defined as an NIHSS score of 0 or 1, at Day 90. A second binary endpoint was clinical improvement, defined as either a complete recovery (inexplicably now defined as an NIHSS score of 0, in contrast to the earlier definition which included also an NIHSS score of 1) or a change from baseline of at least 11 points. If we ignore the inconsistency in how "complete recovery" is defined (first as an NIHSS score of 0 or 1, then as just 0), then clearly a complete recovery implies a clinical improvement, so the two endpoints are fusible, and we really have a single trichotomous endpoint, (no improvement, clinical improvement, complete recovery), where "no improvement" is short hand for either no improvement or improvement not reaching the threshold for clinical improvement. With this endpoint, the data appear to be (we cannot be sure, since only proportions, and not actual patient counts, were presented in the original report) (207, 19, 80) at Day 30 for the placebo arm and (181, 25, 101) at Day 30 for the rt-PA arm. The one-sided Smirnov test yields *p*-values of 0.02094 (exact) and 0.09884 (approximate). At Day 90 the data were (187, 15, 104) in the placebo group and (169, 32, 106) in the rt-PA group, with corresponding one-sided *p*-values of 0.0938 (exact) and 0.3242 (approximate).

Example S8 comes from a study of St. John's wort for major depression. Shelton et al. (2001) measured depression with two binary endpoints, specifically remission and response. Remission is defined as HAM-D ≤ 7 and CGI-I 1 or 2, whereas response is defined as HAM-D ≤ 12 and CGI-I 1 or 2. Clearly these two endpoints are fusible, because a remission implies a response, so the IPCE would be (no response, response without remission, remission), and the data were (83, 14, 5) in the placebo group ($n = 102$) and (72, 12, 14) in the St. John's wort group ($n = 98$). The Smirnov *p*-values were 0.0766 (exact) and 0.4147 (approximate).

Example S9 comes from a study of combination therapy in adults with HIV. Staszewski et al. (2001) presented two binary outcomes, HIV RNA levels of 50 copies per mL or less and HIV RNA levels of 400 copies per mL or less. Obviously, the former implies the latter, so we again have a trichotomous IPCE of fusible endpoints, copies (> 400, 50-400, < 50). What was called the intent-to-treat population was certainly not that, as it excluded 35 of the 562 patients randomized. The true data set, as best as it can be reconstructed from the incomplete presentation published, is (149, 29, 104) in the abacavir arm and (144, 15, 121) in the indinavir arm, each in the presence of lamivudine and zidovudine (hence combination therapy). The Smirnov *p*-values were 0.1051 (exact) and 0.3238 (approximate).

Several recurrent themes emerge from the examples in this section. First, and most obvious, notice that the exact Smirnov test always provides a lower *p*-value than the approximate Smirnov test does, and notice also that in most cases, the approximate one is not even close. It seems reasonable, then, to suggest that the approximate Smirnov test never be used in practice, even if other approximations are accepted.

Examples of the *t*-Test Failing

The *t*-test is often used for continuous outcomes when the variance is not known. It is somewhat ironic that, while we are up front about not knowing the variance, we still wish to cling to this notion that we can somehow know that the data are normally distributed, despite Geary (1947) stating clearly that no data are normally distributed. Table 5 presents four examples in which the *t*-test gave results that differed markedly from corresponding exact results.

Table 5. Data sets for which the *t*-test fails badly

N	References	*p*-values*
T1.	Williams et al. (2000); Barber and Thompson (2000)	0.01, 0.79
T2.	Chaudhry et al. (2002); Jacobs (2003)	0.054, 0.004
T3.	Chaudhry et al. (2002); Jacobs (2003)	0.21, 0.016
T4.	Chaudhry et al. (2002); Jacobs (2003)	0.054, 0.006

Note: Citations abbreviated for space; see Reference section below for full reference

Example T1 bears some similarity to Example C6, in that one set of authors argued that an approximate test should be used after it was already established that an exact method was needed. In this case, the context was open access follow-up for inflammatory bowel disease, and its effect on costs. One particular endpoint was secondary care costs. Williams et al. (2000) correctly pointed out that:

> "Because data on use of resources tend to be highly skewed, routine parametric statistics are not appropriate. We therefore assessed significance by the Mann-Whitney U-test." (p. 545)

Using this proper analysis, the between-group *p*-value for secondary care costs is presented in Table 4 of the original article as 0.01, based on a mean cost of 582 (SD = 808) in the open access arm and 611 (SD = 475) in the routine care arm (the units are not provided in the table). Barber and Thompson (2000) argued that the means are most relevant, and:

> "[T]he most appropriate simple method for comparing mean costs is the ordinary t-test. By using the means and standard deviations in each group reported by the authors, we have calculated *p*-values from *t*-tests ... one of the authors' main conclusions – that open access follow-up used fewer resources in secondary care – is not supported: The *p*-value from the t-test is 0.79." (p. 1730)

Berger (2002) noted that there are two issues here, specifically the choice of test statistic (difference of means, difference of mean ranks, difference of Van der Waerden normal scores, or something entirely different) and the mode of generating a reference distribution. Differences in means can be accompanied by differences in shape and/or spread, so the *t*-test certainly is not always the most powerful test, even to detect the difference in means. But aside from this, even if we were to decide upon the difference of means as the test statistic, this certainly should not imply that we also use an approximation instead of an exact analysis. One can easily conduct an exact *t*-test, using the difference of means as the test statistic, and the permutation reference distribution to evaluate statistical significance.

Examples T2-T4 all come from the same study. Specifically, Chaudhry, Schroter, Smith, and Morris (2002) used the approximate *t*-test for five measures of readers' perceptions of papers with and without declarations of competing interests. These measures were interest, importance, relevance, validity, and believability, and the corresponding *p*-values for the five measures were 0.004, 0.016, 0.006, 0.001, and < 0.001. Jacobs (2003) re-analyzed the data with exact methods, after pointing out the flaws in using approximate methods for the data at hand. Three of the *p*-values became non-significant, specifically interest ($p = 0.054$), importance ($p = 0.21$), and relevance ($p = 0.054$). Of course, 0.054 is close to 0.05, so one might be tempted to declare it close enough. This is bad policy, and bad statistics, and not to be confused with selecting an alpha level other than 0.05. While it is perfectly reasonable to select an alpha level other than 0.05, maybe even 0.055, this selection needs to be made prior to viewing the data (and the *p*-value). Otherwise, one is left wondering just how broad this fuzzy

inclusion region actually is. Would 0.06 have been OK? What about 0.07? Where is the line drawn? In other words, what is alpha? And if alpha is not what we said it was up front, then we have a problem with the usage of alpha, and we are drawing the bull's eye around where the dart happened to hit.

Moreover, notice that the *p*-value for importance went from 0.016 to 0.21 when the analysis went from approximate to exact. This, as well as some of the other examples in Table 1, may well surprise those who consider the choice of an exact or an approximate test to be a "fourth decimal problem" that hardly warrants the attention of today's modern statistician. The StatXact manual states that "It is wise to never report an asymptotical p-value without first checking its accuracy against the corresponding exact or Monte Carlo *p*-value. One cannot easily predict a priori when the asymptotic *p*-value will be sufficiently accurate" (Cytel Software, 1995, p. 21). This is certainly excellent advice, but we can go a step further and ask why one would then discard the gold standard, the exact permutation *p*-value, once it is in hand, to use instead an approximation to it?

Summary and Conclusions

"Robustness procedures are generally considered to be statistical methods which are insensitive to small deviations from the underlying assumptions" (Prescott, 1998, p. 3864), and often this vagueness regarding how insensitive and how small the deviations must be allows for excessive discretion in filling in the blanks. That is to say that many researchers operate as if this robustness is absolute, so that there is no sensitivity at all no matter the magnitude of the deviation or how it is quantified. In point of fact, there seems to be no reliable method for imputing an exact *p*-value based on only the combination of knowledge of the approximate *p*-value and appeal to this alleged robustness. The fact that an exact *p*-value can fall anywhere on the unit interval even once we know the value of the approximate *p*-value should serve as ample demonstration that any notion of robustness being absolute is an illusion.

There might still be a value in computing approximate *p*-values anyway, if there were some added cost or difficulty involved in computing the exact *p*-value. In some applications this in fact is the case, but certainly not in all, and it is worth the effort to determine which case we are in. If an exact *p*-value can be computed relatively easily, with no prohibitive cost, then it is difficult to imagine any valid argument for not doing so. This remains the case even if one can put forth a compelling argument in favor of presenting an approximate *p*-value. For example, it may be the case that precedent favors the approximate *p*-value, which has

always been computed in the past. We want to see how the present data compare to past data sets, and those older ones were summarized, for example, with *t*-tests, and we do not have access to the complete data that would enable us to conduct exact analyses of those older data sets. In this case, it seems reasonable to compute the *t*-test on the new data set for the sake of comparing apples to apples and oranges to oranges, but this does not preclude the possibility of also computing an exact *p*-value in addition to the approximate one. Under no circumstances should we ever pretend to know the exact *p*-value without actually computing it.

Acknowledgements

The review team offered helpful comments that resulted in a far improved final version.

References

Barber, J. A., & Thompson, S. G. (2000). Would have been better to use t-test than Mann Whitney U-test. *BMJ, 320*, 1730. doi: 10.1136/bmj.320.7251.1730

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine*, *19*(10), 1319-1328. doi: 10.1002/(sici)1097-0258(20000530)19:10<1319::aid-sim490>3.0.co;2-0

Berger, V. W. (2002). Improving the information content of categorical clinical trial endpoints. *Controlled Clinical Trials*, *23*(5), 502-514. doi: 10.1016/s0197-2456(02)00233-7

Berger, V. W. (2004). On the generation and ownership of alpha in medical studies. *Controlled Clinical Trials*, *25*(6), 613-619. doi: 10.1016/j.cct.2004.07.006

Berger, V. W., & Lachenbruch, P. A. (1998). *Robust permutation tests and randomized clinical trials*. Unpublished internal document, United States Food and Drug Administration.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Upper Saddle River, NJ: Prentice Hall.

Chaudhry, S., Schroter, S., Smith, R., & Morris, J. (2002). Does declaration of competing interests affect readers' perceptions? A randomized trial. *BMJ*, *325*, 1391-1392. doi: 10.1136/bmj.325.7377.1391

Clancy, W. G. (2000). Letter to the editor. *American Journal of Sports Medicine, 28*(4), 615.

Clark, W. M., Wissman, S., Albers, G. W., Jhamandas, J. H., Madden, K. P., & Hamilton, S. (1999). Recombinant tissue-type plasminogen activator (Alteplase) for ischemic stroke 3 to 5 hours after symptom onset: The ATLANTIS study: A randomized controlled trial. *JAMA: Journal of the American Medical Association, 282*(21), 2019-2026. doi: 10.1001/jama.282.21.2019

Cytel Software. (1995). *StatXact 3 for Windows: Statistical software for exact nonparametric inference: User manual*. Cambridge, MA: Cytel Software Corporation.

Elwood, J. M. (1998). *Critical appraisal of epidemiological studies and clinical trials* (2nd ed.). Oxford, UK: Oxford University Press.

Fentiman, I. S., Rubens, R. D., & Hayward, J. L. (1983). Control of pleural effusions in patients with breast cancer: A randomized trial. *Cancer, 52*(4), 737-739. doi: 10.1002/1097-0142(19830815)52:4<737::AID-CNCR2820520428>3.0.CO;2-8

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.

Fox, S. M., Einhorn, L. H., Cox, E., Powell, N., & Abdy, A. (1993). Ondansetron versus ondansetron, dexamethasone, and chlorpromazine in the prevention of nausea and vomiting associated with multiple-day cisplatin chemotherapy. *Journal of Clinical Oncology, 11*(12), 2391-2395. doi: 10.1200/jco.1993.11.12.2391

Geary, R. C. (1947). Testing for normality. *Biometrika, 34*(3/4), 209-242. doi: 10.2307/2332434

Hewett, T. E., Lindenfeld, T. N., Riccobene, J. V., & Noyes, F. R. (1999). The effect of neuromuscular training on the incidence of knee injury in female athletes. *The American Journal of Sports Medicine, 27*(6), 699-706.

Hewett, T. E., Levy, M., Lindenfeld, T. N., & Noyes, F. R. (2000). Author's response. *The American Journal of Sports Medicine 28*(4), 615-616.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York, NY: Wiley.

Jacobs, A. (2003). Clarification needed about possible bias and statistical testing. *BMJ USA, 3*, 93.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco, CA: Holden Day.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician, 43*(4), 283-288. doi: 10.2307/2685390

Lu, J., Ding, P., & Dasgupta, T. (2015). Construction of alternative hypotheses for randomization tests with ordinal outcomes. *Statistics & Probability Letters, 107*, 348-355. doi: 10.1016/j.spl.2015.09.013

Prescott, P. (1998). Robustness. In *The Encyclopedia of Biostatistics*. (Vol. 5, pp. 3864-3869). Chichester, UK: Wiley.

The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment Investigators. (1998). Low molecular weight heparinoid, ORG 10172 (danaparoid), and outcome after acute ischemic stroke: A randomized controlled trial. *JAMA: Journal of the American Medical Association, 279*(16), 1265-1272. doi: 10.1001/jama.279.16.1265

Rigdon, J, & Hudgens, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine, 34*(6), 924-935. doi: 10.1002/sim.6384

Shelton, R. C., Keller, M. B., Gelenberg, A., Dunner, D. L., Hirschfeld, R., Thase, M. E.,… Halbreich, U. (2001). Effectiveness of St John's wort in major depression: A randomized controlled trial. *JAMA: Journal of the American Medical Association, 285*(15), 1978-1986. doi: 10.1001/jama.285.15.1978

Staszewski, S., Keiser, P., Montaner, J., Raffi, F., Gathe, J., Brotas, V.,… Spreen, W. (2001). Abacavir-Lamivudine-Zidovudine vs Indinavir-Lamivudine-Zidovudine in antiretroviral-naive HIV-infected adults: A randomized equivalence trial. *JAMA: Journal of the American Medical Association, 285*(9), 1155-1163. doi: 10.1001/jama.285.9.1155

Williams, J. G., Cheung, W. Y., Russell, I. T., Cohen, D. R., Longo, M., & Lervy, B. (2000). Open access follow-up for inflammatory bowel disease: Pragmatic randomized trial and cost effectiveness study. *BMJ, 320*, 544-548. doi : 10.1136/bmj.320.7234.544

Zelterman, D., Chan, I. S. F., & Mielke, P. W. (1995). Exact tests of significance in higher dimensional tables. *The American Statistician, 49*(4), 357-361. doi: 10.2307/2684573

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 51.

ISSN 1538 – 9472

Regular Articles

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 52-68.
doi: 10.22237/jmasm/1493596980

ISSN 1538 − 9472

Experiment-wise Type I Error Rates in Nested (Hierarchical) Study Designs

Jack Sawilowsky
Citigroup
Tampa, FL

Barry Markman
Wayne State University
Detroit, MI

When conducting a statistical test one of the initial risks that must be considered is a Type I error, also known as a false positive. The Type I error rate is set by nominal alpha, assuming all underlying conditions of the statistic are met. Experiment-wise Type I error inflation occurs when multiple tests are conducted overall for a single experiment. There is a growing trend in the social and behavioral sciences utilizing nested designs. A Monte Carlo study was conducted using a two-layer design. Five theoretical distributions and four real datasets taken from Micceri (1989) were used, each with five different sample sizes and conducted with nominal alpha set to 0.05 and 0.01. These were conducted both unconditionally and conditionally. All permutations were performed for 1,000,000 repetitions. It was found that when conducted unconditionally, the experiment-wise Type I error rate increases from alpha = 0.05 to 0.10 and 0.01 increases to 0.02. Conditionally, it is extremely unlikely to ever find results for the factor, as it requires a statistically significant nest as a precursor, which leads to extremely reduced power. Hence, caution should be used when interpreting nested designs.

Keywords: Experiment-wise Type I error inflation, nested testing, Monte Carlo simulation, hierarchical linear modeling, Bonferroni-Dunn

Type I Error

When conducting a statistical test one of the initial risks that must be considered is a Type I error, also known as a false positive. It occurs by "rejecting a null hypothesis when it is true" (Hinkle, Wiersma, & Jurs, 2003, p. 178). It is set by nominal alpha, assuming all underlying conditions of the statistic are met. For example, if nominal $\alpha = 0.05$, then this indicates the threshold for what constitutes a rare event is set to a 5% probability of a false positive, or odds corresponding to less than or equal to 1 in 20.

Jack Sawilowsky, Ph.D,. is a Vice President – Senior Data Scientist. Email him at: jsitm585@gmail.com. Barry Markman is a Professor of Educational Psychology. Email him at: barry.markman@wayne.edu.

The risk represented by the Type I error only applies if a single statistical test is conducted on the data set. If multiple analyses are conducted, the Type I error rate will increase above nominal alpha. This is known as experiment-wise Type I error inflation: the "Experimentwise error rate (α_E) is the probability of making a Type I error rate for the set of all possible comparisons" (Hinkle et al., 2003, p. 372). Statisticians have considered this problem since the second half of the 20th Century and have proposed a variety of solution strategies to handle Type I error inflation, particularly for statistical approaches that invoke multiple procedures.

Type I error inflation can arise in many statistical procedures. In some circumstances, such as the one-way independent samples ANOVA layout, there is a storied history of the development of a priori and post-hoc corrections to the *F* test to ameliorate this problem. Unfortunately, the experiment-wise inflation problem does surface in certain seemingly innocuous layouts, and results are often presented without recognizing the need for adjustment.

According to some viewpoints, there are also statistical layouts that permit a step-down analysis. An example is following a multivariate test (e.g., MANOVA or MANCOVA) with univariate tests. Consider a Hotellings' T^2 which conceptually is an extension of the test of difference in means in the Student's *t* test to the multivariate case, which is the difference in group centroids. A question that frequently arises following a significant T^2 is if one or the other dependent variable was the greater contributor.

Suppose both a test of reading and mathematics achievement were given following an intervention, and the T^2 test of differences in means between females and males was statistically significant. The step-down univariate test (i.e., Student's *t* test) on reading by gender, and mathematics by gender, would then be conducted. The statistical literature is not settled on the appropriateness of this approach. The general consensus is if the multivariate test was conducted only to maximize power there is no reason why step-down tests shouldn't be conducted (other than the inflation of Type I errors). However, if the T^2 was conducted because of a multivariate hypothesis with intertwined dependent variables (e.g., self-esteem and self-worth), conducting step-down tests and the concern with experiment-wise Type I error inflation vanishes.

There are, however, other layouts that according to all viewpoints require multiple statistical tests. The classical example of this is the one-way analysis of variance. The omnibus *F* test can be used to determine if there is a difference in means somewhere within the $K \geq 3$ groups. Either a priori or post-hoc comparisons must be conducted in order to determine precisely where the

difference(s) in means occurred. It is recognized that conducting multiple tests in this application increases the experiment-wise Type I error rate.

Sequential (or Serial) Tests

Sequential tests occur in separate phases. For example, there is the recommendation to test for underlying assumptions (e.g., homoscedasticity via Levine's test and normality via Kolmogorov-Smirnov's test), and only after failing to reject both proceeding to conduct a statistical test of effects (such as the *t*-test). This strategy was recommended in many statistical packages (e.g., Statistical Analysis Systems Institute, Inc., 1990, p. 25; Norušis, 1993, pp. 254-255; Wilkinson, 1990, p. 487). However, Sawilowsky (2002) noted, "There is a serious problem with this approach that is universally overlooked. The sequential nature of testing for homogeneity of variance as a condition of conducting the independent samples *t*-test leads to an inflation of experiment-wise Type I errors" (p. 466). Sawilowsky (2002) conducted a Monte Carlo study that demonstrated the experiment-wise Type I error rate inflated to almost twice alpha. A possible solution to this is to avoid using a parametric test that requires testing for underlying assumptions when the data are not known to be normally distributed and homogeneous, and using a nonparametric alternative in its place.

Parallel Tests

Parallel tests occur when multiple tests are conducted at the same time. For example, in ANOVA, multiple main effects and interactions can all be of interest. There is debate whether to start with the main effects or interactions, and whether to stop or continue after finding significance (see, e.g., Sawilowsky, 2007a, ch. 14). Regardless of the method chosen, all tests are conducted simultaneously. For example, with three main effects, the following seven combinations can be tested for significance: A × B × C, A × B, A × C, B × C, A, B, and C.

There is a commonly held belief by researchers that ANOVA provides weak protection against the inflation of Type I error rates when conducting multiple tests. This is due to the researcher being genuinely interested in multiple hypotheses. It is believed that this interest adequately negates the effect of conducting repeated measures while utilizing the Frequentist approach. It is argued that ANOVA is in contrast to processes such as stepwise regression, in which the researcher does not have prior suspicion or even interest in the various hypotheses being tested. However, Kromrey and Dickenson (1995) stated:

> In a two-factor ANOVA, three null hypotheses are tested (one for each main effect and one for the interaction effect), while in a three-factor analysis, seven null hypotheses are tested (three main effects, three first-order interactions, and one second-order interaction), and in a four-factor analysis, fifteen null hypotheses are tested. The effects of multiple testing… in factorial ANOVA has not been undertaken, despite the fact that the problem has been recognized for more than 30 years. (pp. 51-52)

They conducted a Monte Carlo simulation in which the number of factors (2-4), pattern of effects (null and/or non-null), effect size (small-large), and sample size (5, 10, and 20) were modeled. The simulation was conducted with 5,000 repetitions per experimental condition. In order to safeguard against rival hypotheses affecting the results, the ANOVA *F* tests were conducted on data sampled from a theoretical normal distribution, thus ensuring internal validity.

Conditioned on a significant omnibus *F* test, with the two-factor model, the experiment-wise Type I error rate for the null effects were 0.06. With the three-factor model, it was as high as 0.16, and with four factors, it rose to 0.35 for the null effects. These results demonstrated that the issue of experiment-wise Type I error rate applies to the parallel scenario, even in the presence of a known significant non-null effect. In other words, the weak protection is ineffective in controlling experiment-wise Type I error rate inflation.

Post-Hoc Tests: A Resolution to the Type I Error Inflation Problem

Wilcox (1996) described the most extreme post hoc solution to experiment-wise Type I error inflation:

> The Bonferroni procedure, sometimes called Dunn's Test, provides a simple method of performing two or more tests such that the experimentwise Type I error probability will not exceed α. If you want experimentwise Type I error probability to be at most α, you simply perform paired *t*-tests, each at the $\alpha b = \alpha/C$ level of significance, where *C* is the total number of comparisons you plan to perform. (pp. 279-280)

The Bonferroni-Dunn procedure divides alpha by the number of tests to be conducted, to ensure that after all hypothesis tests are computed the total Type I

error rate does not exceed nominal alpha. This method is guaranteed to contain the Type I error rate, but it also guarantees loss of statistical power, because as α decreases, β increases; and as β increases, power decreases (Hinkle et al., 2003, p. 300). All other multiple comparison procedures are a compromise between the Bonferroni and making no adjustments to control Type I error inflations.

Nesting

Hierarchical linear modeling (HLM), which is based on testing nested effects, is a popular statistical approach to school-based research. Kreft and De Leeuw (1998) stated, "Hierarchical data structures are very common in the social and behavioral sciences… Once you know that hierarchies exist, you see them everywhere" (p. 1). Kanji (1999) provided a definition of a nested or hierarchical classification as follows:

> In the case of a nested classification, the levels of factor B will be said to be nested with the levels of factor A if any level of B occurs with only a single level of A. This means that if A has p levels, then the q levels of B will be grouped into p mutually exclusive and exhaustive groups, such that the i^{th} group of levels of A is q_i, i.e. we consider the case where there are $\sum_i q_i$ levels of B. (p. 128)

Winer (1971) explained, "Effects which are restricted to a single level of a factor are said to be nested within that factor" (p. 360). Winer emphasized the substantial limitation of nested designs in that they do not permit the testing of an interaction effect.

As an example of a nested design, consider a teacher within school layout. Kanji (1999) decomposed the three components (A School factor, B Teacher factor, Residual) nested sums of squares as

$$S_{\text{S}}^2 = \sum_i n_i \left(Y_{i00} - Y_{000}\right)^2,$$

$$S_{\text{T}}^2 = \sum_i \sum_j n_{ij} \left(Y_{ij0} - Y_{i00}\right)^2, \text{ and}$$

$$S_{\text{E}}^2 = \sum_i \sum_j \sum_k \left(Y_{ijk} - Y_{ij0}\right)^2$$

Table 1. Nested design example data from Kanji (1999, p. 129)

	Schools											
	I			II			III			IV		
	Teacher			Teacher			Teacher			Teacher		
	1	2	3	1	2	3	1	2	3	1	2	3
	44	39	39	51	48	44	46	45	43	42	45	39
	41	37	36	49	43	43	43	40	41	39	40	38
	39	35	33	45	42	42	41	38	39	38	37	35
	36	35	31	44	40	39	40	38	37	36	37	35
	35	34	28	40	37	37	36	35	34	34	32	35
	32	30	26	40	34	36	34	34	33	31	32	29
TT	227	210	193	269	244	241	240	230	227	220	223	211
$\bar{X}_T$	37.80	35.00	32.17	44.83	40.67	40.16	40.00	38.33	37.83	36.67	37.17	35.17
ST	630			754			679			654		
$\bar{X}_S$	35			41.89			38.72			36.33		

Note: TT = Teacher total, ST = School total, $\bar{X}_T$ = Teacher mean, $\bar{X}_S$ = School mean, Grand mean School total = 2,735

Table 2. Kanji (1999, p. 130) ANOVA table

	df	SS	Mean Square	*F*
Schools	3	493.60	164.53	6.47
Teachers within School	8	203.55	25.44	1.46
Pupils within Teachers	60	1047.84	17.46	
Total	71	1744.99		

where S is the School, T is the Teacher, and E is the residual, where H_A: $\alpha_i = 0$ for all i and H_B: $\beta_{ij} = 0$ for all i, j. The data for the example are compiled in Table 1, and the traditional ANOVA table is presented in Table 2.

Hierarchical Modeling

Kreft and De Leeuw (1998) stated that hierarchical modeling tends to address research questions that lack independence and other experimental conditions, which makes it incompatible with ANCOVA (p. 5). Similarly, Kennedy and Bush (1985) noted "Interaction is not a meaningful consideration when one variable is nested within another" (p. 52). For an interaction effect to be measured, all factors in all levels would need to contain all factors of all other levels. However, nesting is advantageous in order to control for unique effects of a specific level of a nest on another level (e.g., schools on curriculum).

There are also more sophisticated multi-level and longitudinal models based on these basic layouts (Heck, Thomas, & Tabata, 2010). However, there has been little discussion in the literature regarding the impact on the inflation of experiment-wise Type I error rates due to the hierarchical testing of treatment effects. For example, Kanji (1999) did not address the issue of conducting multiple *F* tests following the results obtained in Table 2 above. If each test is set at $\alpha = 0.05$, then in reality there will be an approximate experiment-wise Type I error rate of 0.10. Similarly, Winer's (1971) presentation of the different types of nested designs (2 Factors, Partial, and 3 or more Factors) was not accompanied by a discussion on the experiment-wise Type I error rate.

Methodology

Design

A two-factor nested layout or hierarchical classification layout was used. This design assumed errors would be normally distributed, with the magnitudes of

those errors being independent from either of the two factors. Specifically, the hypothetical layout pertained to an analysis of difference of means between classes taught by different teachers, with teachers in turn being nested within different schools. In this layout, student test scores were simulated for three teachers (or classrooms) per each of four schools, as noted in the table below.

Nested designs are almost always conducted through the use of multiple ANOVA tests. Others, such as the *t* test, are generally not found, because rarely are such studies conducted on two schools with two teachers per school (e.g., Kanji, 1999; Winer, 1971). Therefore, when a nested layout is found in the literature, generally the ANOVA test is required.

Sampling Plan

A pseudo-random number generator was used to simulate student test scores. The data were generated through Roguewave's (2012) subroutine libraries for the theoretical distributions. Data were simulated to follow the Gaussian, uniform, exponential, *t* (df = 3), and Chi-squared (df = 2) distributions. Variates from the Gaussian (i.e., normal) distribution were used to demonstrate the veracity of the Fortran coding. Deviates from non-normal distributions are commonly used in Monte Carlo studies to illustrate robustness properties with respect to Type I errors for departure from population normality.

Samples were also obtained from real data sets (Micceri, 1989) via the Realpops 2.0 subroutine library (Sawilowsky & Fahoome, 2003); Realpops 2.0 is a Fortran 90 updated version of the Fortran 77 subroutine library by Sawilowsky, Blair, and Micceri (1990). For details on the real data sets, see Micceri (1989) and Sawilowsky and Blair (1992). The real data sets to be sampled were the smooth symmetric (achievement scores), digit preference (achievement scores), multi-modal lumpy (achievement scores), and extreme asymmetry (psychometric scores).

Sample sizes were set to n = 2, 10, 30, 45, and 120. Samples of size n = 2 and n = 120 were selected to represent the theoretical minimum and a reasonable maximum study parameter, as is customarily done in Monte Carlo studies. Samples of size n = 10, 30, and 45 were selected to represent small, medium and large classrooms, respectively. Under the truth of the null hypothesis (and homoscedasticity as modeled in this study), unbalanced layouts (i.e., unequal sample sizes per teacher or unequal teachers per school) have no impact on Type I errors and are therefore not modeled. One million repetitions were executed for each combination of study parameters.

Table 3. Expected Type I error rates for normal and selected non-normal data at $\alpha = 0.05$ and $\alpha = 0.01$

Distribution / Dataset	Resulting alpha (0.05)	Resulting alpha (0.01)
Normal	0.050	0.010
Exponential[1]	0.040	0.004
Uniform[1]	0.051	0.010
Digit preference[2]	0.050	0.012
Extreme asymmetric[2]	0.047	0.009
Multi-modal lumpy[2]	0.052	0.012
Smooth symmetric[2]	0.050	0.010

Note: [1]Glass, Peckham, and Sanders (1972, p. 250); [2]Sawilowsky and Blair (1992, pp. 356-358); these results are for different numbers of repetitions and are based generally on the balanced layout of samples sizes $n_1 = n_2 = 20$; increasing the number of repetitions and sample sizes will give Type I errors closer to nominal alpha

Analysis

The appropriate analysis for the nested design in Table 1 above is a series of two *F* tests. Initially, the *F* test was conducted to determine if there are teacher differences. Under ideal conditions, the intent is to fail to reject the null hypothesis. This is because it is assumed that the teachers have similar qualifications (e.g., certification, experience) in order to be named the instructor of record.

The more important test was then conducted. This is an *F* test for effects, which in this case is for the difference in means between schools. When the null hypothesis was false, it meant the new curriculum administered in at least one school statistically significantly changed student scores. The *F* test should reject this null hypothesis.

In the current study, the truth of the null hypothesis is based on the generation of pseudo-random numbers. There was an expected Type I error rate for each of the component tests. The experiment-wise Type I error rate will be determined by the sum of those two Type I error rates.

This will be accomplished in two ways. The first is unconditional; meaning the test for effects (i.e., between schools) will be conducted regardless of the results of the test for nesting (i.e., between teachers). The second is conditional; meaning the test for effects will only be conducted if and only if a nesting effect is non-null.

Differentiating between unconditional and conditional testing is advisable if the general purpose for conducting an intervention study is to determine if there is a difference between schools where students did or did not receive an intervention.

The impact of teacher differences should be negligible. In other words, the school effect should only be tested when it can be first shown there was no teacher effect.

In order to increase generality of results, the *F* tests invoked in the Monte Carlo simulation were conducted at both the nominal $\alpha = 0.05$ and 0.01 levels.

Error Isolation

The Monte Carlo simulation was conducted using parametric or normal theory tests. However, data were also drawn from non-normal distributions. Therefore, the issue arises as to where potential results are originating. If the Type I error rates do inflate, it is important to determine whether these results are due to experiment-wise Type I error inflation or if they are caused by violating the assumption of normality. Typical Type I error rates are listed in Table 3.

Results

Unconditional

The test for the nest and the treatment effect are both conducted in this model of analysis. Although it does not matter which test is conducted first, for consistency, the test for the nest was conducted prior to the test of the effect. A series of tabled results are presented, arranged by distribution or dataset type. The entries inside each table represent the Type I error rate for the study conditions.

As predicted by theory (Marascuilo & Serlin, 1988), the results in Tables 4 and 5 demonstrate that conducting a series of two statistical tests unconditionally, regardless of the nature of those tests, produces an experiment-wise Type I error rate of approximately twice nominal alpha. Tables 4 and 5 contain a compilation of those results.

In Tables 6 and 7, the Type I error rates are averaged as in the previous two tables, except the test for the factor (i.e., School) is conducted conditionally subsequent to a significant test of the nesting effect. In order to understand these results, consider Bradley's (1968) definition for two levels of robustness. The conservative definition is met when the Type I error rate is within the bounded interval $[0.5\alpha, 1.5\alpha]$ inclusive, and the liberal definition is met when the Type I error rate is within the bounded interval $[0.9\alpha, 1.1\alpha]$ inclusive. The results for the factor (School) are ultra-conservative, falling far below 0.025 when the test is conducted at the 0.05 nominal alpha level, and below 0.005 when the test is conducted at the 0.01 nominal alpha level. In addition, the impact of being ultra conservative means the test for the factor (School) greatly lacks statistical power.

Table 4. Summary of average Type I error rates for various distributions/datasets, unconditional, $\alpha = 0.05$

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.050039	0.050070	0.100109
Chi-square (df=3)	0.050073	0.049391	0.099464
Exponential	0.050012	0.049008	0.099019
t (df=3)	0.045460	0.045810	0.091269
Uniform	0.051215	0.050653	0.101868
Digit preference	0.050246	0.050201	0.100446
Extreme asymmetric	0.052485	0.050207	0.102693
Multi-modal lumpy	0.052758	0.050786	0.103544
Smooth symmetric	0.050241	0.050236	0.100477

Table 5. Summary of average Type I error rates for various distributions/datasets, unconditional, $\alpha = 0.01$

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.010042	0.010006	0.020048
Chi-square (df=3)	0.010618	0.010236	0.020854
Exponential	0.011089	0.010254	0.021343
t (df=3)	0.008624	0.008728	0.017353
Uniform	0.010595	0.010286	0.020881
Digit preference	0.010117	0.010093	0.020210
Extreme asymmetric	0.012795	0.011150	0.023944
Multi-modal lumpy	0.011357	0.010315	0.021672
Smooth symmetric	0.010106	0.010142	0.020247

Table 6. Summary of average Type I error rates for various distributions/datasets, conditional, $\alpha = 0.05$

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.050039	*0.000357*	0.050397
Chi-square (df=3)	0.050073	*0.000472*	0.050545
Exponential	0.050012	*0.000489*	0.050500
t (df=3)	0.045460	*0.000304*	0.045763
Uniform	0.051215	*0.000563*	0.051777
Digit preference	0.050246	*0.000425*	0.050671
Extreme asymmetric	0.052485	*0.000770*	0.053256
Multi-modal lumpy	0.052758	*0.000609*	0.053367
Smooth symmetric	0.050241	*0.000411*	0.050652

Note: Values in italics are nonrobust according to Bradley's (1968) liberal definition

Table 7. Summary of average Type I error rates for various distributions/datasets, conditional, $\alpha = 0.01$

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.010042	*0.000020*	0.010062
Chi-square (df=3)	0.010618	*0.000014*	0.010632
Exponential	0.011089	*0.000012*	0.011101
t (df=3)	0.008624	*0.000000*	0.008624
Uniform	0.010595	*0.000016*	0.010612
Digit preference	0.010117	*0.000000*	0.010117
Extreme asymmetric	0.012795	*0.000050*	0.012845
Multi-modal lumpy	0.011357	*0.000000*	0.011357
Smooth symmetric	0.010106	*0.000000*	0.010106

Note: Values in italics are nonrobust according to Bradley's (1968) liberal definition

Statistical Power Projections

As previously noted, conducting the test of the factor (i.e., School) conditionally will create a lack of statistical power due to the ultra-conservative nature of being the second in sequence in a series of two tests. Although it is beyond the scope of the current study to conduct a full-scale power spectrum analysis, in an attempt to explain the impact on statistical power, a treatment alternative of shift in location parameter was introduced.

The study parameters for this brief power study included setting nominal $\alpha = 0.05$. Data were sampled from the Gaussian distribution, the sample size was set at $n = 2$, and both unconditional and conditional testing were conducted. The treatment was modeled by the addition of a constant equal to 0.5σ, where $\sigma = 1$ when the referent distribution is normal, to create an effect size of Cohen's $d = 0.5$. The magnitude of this effect size is considered moderate (Cohen, 1988).

The treatment conditions were set in two studies as follows. For Study 1, an effect size of 0.5 was added to a single teacher per school. This created a difference among the twelve teachers, while leaving the schools equal. For Study 2, all teachers in a single school were simulated to receive the treatment, creating a difference between both the teachers and the schools. Due to the layout of nested designs, in this case with teachers contained within the school where they work, it is impossible to simulate a change between schools only. The results are compiled in Table 8.

As noted, with the given study parameters, the unconditional and conditional power for the test of the nest effect (Teacher) was 0.194. In the unconditional layout, the expected Type I error rate of approximately 0.05 was

obtained; however, in the conditional, the Type I error rate was ultra-conservative at 0.011. The loss in power becomes apparent in Study 2. Although the power was approximately the same for the treatment effect (0.121 and 0.114, respectively) for the conditional layout, the power obtained for the effect (school) was reduced to from 0.141 to 0.089, which is a severe loss in power of approximately 22%.

Restating and expanding on Kreft and De Leeuw (1998):

> Hierarchical data structures are very common in the social and behavioral sciences… Once you know that hierarchies exist, you see them everywhere… Examples include students nested within schools, employees nested within firms, or repeated measurements nested within persons. (p. 1)

Similarly, Gonzales (2009) indicated when the "factors are not crossed… we cannot use the machinery of the factorial analysis of variance" (p. 313). The proposed solution is to turn to nested designs, which are "now a major area of research in social science statistics" (p. 314). Gonzales concluded: "Multilevel modeling techniques permit simultaneous modeling of all the levels that are accounted for in the design" (p. 315).

Unfortunately, the observations of Kreft and De Leeuw and Gonzales overlooked the impact of conducting statistical tests in a hierarchical model in general and in nested designs in particular. Gonzales (2009) attempted to forestall the impact of multiple testing with the rhetorical question, "Aren't we capitalizing on chance by making so many comparisons?" (p. 336). The first answer given was to make nested designs analogous to factorial ANOVA where there appears to be no concern in the statistical literature over the inflation of Type I error in testing main effects and interactions. However, as noted by Kromrey and Dickenson (1995), and discussed at length earlier in this article, this provides no safe haven from experiment-wise Type I error inflation.

Table 8. Statistical power projections, normal distribution, $\alpha = 0.05$, $n = 2$

				Power			
	Study Parameters			Unconditional		Conditional	
Recipeint	a	ES Teacher	ES School	Teacher	School	Teacher	School
Teacher	0.05	0.5	0.0	0.194	0.054	0.194	0.011
Teacher and School	0.05	S1 = 0.5	S2-4 = 0.0	0.121	0.114	0.121	0.089

Note: ES = effect size in standard deviations, S1 = School 1, S2-4 = Schools 2, 3, and 4

The second argument advanced by Gonzales (2009) to preclude issues of multiple testing in nested designs was, "Replication is the best way to deal with concerns about multiple tests and inflated Type I error rates" (p. 337). However, Sawilowsky (2007b) demonstrated in a Monte Carlo experiment that "replicating the same poor design has little chance of contributing accurate evidence for or against the effectiveness of a treatment, or for quantifying the magnitude of its effectiveness if it exists" (pp. 221-222).

The third argument advanced by Gonzales (2009) was to apply a correction such as the Bonferroni-Dunn technique (p. 285). This is precisely the solution strategy previously proposed by Kromrey and Dickenson (1995). However, such methods always result in a reduction of statistical power and should be used as a last resort.

Indeed, despite offering these three solution strategies, Gonzales (2009) concluded that experiment-wise Type I error rate inflation was something that researchers need not take seriously. However, to his credit, Gonzales' final word on this issue was "We admit that we are in the minority among methodologists on this particular point" (p. 285).

Hence, the purpose of this study was to explicate the impact of simple nesting designs on experiment-wise Type I error rates via a Monte Carlo exercise. Study parameters included popular population distributions and vetted large datasets to generate samples using common sample sizes and alpha levels for the single nested layout of three teachers per school for four schools. The tests for the nest and effect were conducted unconditionally and conditionally.

Conclusion

Prior to drawing a conclusion in resolving the issue of the impact of nesting on the inflation of experiment-wise Type I error rates, it should be mentioned that there are potentially other statistical techniques that could have been incorporated, such as the nonparametric Kruskal-Wallis and the rank transform tests. Neither test is a solution for the inflation of experiment-wise Type I errors, but it is not known if either would help recover some of the lost power. However, because neither the Kruskal-Wallis nor the rank transform tests have been developed specifically for nested layouts, they were not incorporated in the study.

As Kromrey and Dickenson (1995) showed, the testing of multiple effects in a layout can be safely carried out via invoking a Bonferroni-Dunn or similar technique. However, as it stands, the statistical power available to the testing of the treatment effect conditional on a significant nested effect is already severely

reduced due to the procedure being ultra-conservative. The use of Bonferroni-Dunn or related methods will only further reduce statistical power. When the same issue arose in analyzing the Solomon four-group design (Sawilowsky & Markman, 1990a, b; Sawilowsky, Kelley, Blair, & Markman, 1994), a solution based on an asymmetric Bonferroni-Dunn (i.e., disproportionate allocation of nominal alpha to constituent tests) was proposed by Sawilowsky (1996).

Nevertheless, Heck et al. (2010) noted more sophisticated nested designs "are rapidly growing in their popularity and use" (p. 320), which will only exacerbate the issues outlined in this study. Hence, researchers should heavily weigh the trade-offs of experiment-wise Type I error inflation for unconditional and statistical power loss for conditional nested designs before utilizing them.

References

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, J. (1988). *Power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.2307/1169991

Gonzales, R. (2009). *Data analysis for experimental design*. New York, NY: Guilford Press.

Heck, R. H., Thomas, S. L, & Tabata, L. N. (2010). *Multilevel and longitudinal modeling with IBM SPSS*. New York, NY: Routledge/Taylor & Francis. doi: 10.4324/9780203855263

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences*. New York, NY: Houghton Mifflin.

Kanji, G. K. (1999). *100 statistical tests*. London, UK: Sage. doi: 10.4135/9781849208499

Kennedy, J. J., & Bush, A. J. (1985). *An introduction to the design and analysis of experiments in behavioral research*. Lanham, MD: University Press of America.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage. doi: 10.4135/9781849209366

Kromrey, J. D., & Dickenson, W. B. (1995). The use of an overall *F* test to control Type I error rates in factorial analyses of variance: Limitations and better strategies analyses of variance: limitations and better strategies. *Journal of Applied Behavioral Science, 31*(1), 51-64. doi: 10.1177/0021886395311006

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York, NY: W. H. Freedman and Company.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

Norušis, M. J. (1993). *SPSS for Windows: Base system user's guide, release 6.0*. Chicago, IL: SPSS Inc.

Sawilowsky, S. S. (1996, June 23). *Controlling experiment-wise Type I error in the Solomon four-group design*. Presented at the 1st International Conference on Multiple Comparisons. Tel Aviv, Israel.

Sawilowsky, S. S. (2002). The probable difference between two means when $\sigma_1 \neq \sigma_2$. *Journal of Modern Applied Statistical Methods, 1*(2), 461-472. doi: 10.22237/jmasm/1036109940

Sawilowsky, S. S. (2007a). ANOVA: Effect sizes, interaction vs. main effects, and a modified ANOVA table. In S. S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 191-212). Washington, DC: InfoAge Publishing.

Sawilowsky, S. S. (2007b). ANCOVA and quasi-experimental design: The legacy of Campbell and Stanley. In S. S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 213-238). Washington, DC: InfoAge Publishing.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360. doi: 10.1037/0033-2909.111.2.352

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). REALPOPS.LIB: A PC Fortran library of eight real distributions in psychology and education. *Psychometrika, 55*(4), 729.

Sawilowsky, S. S., & Fahoome, G. F. (2003). *Statistics via Monte Carlo simulation with Fortran*. Rochester Hills, MI: JMASM.

Sawilowsky, S. S., Kelley, D. L., Blair, R. C., & Markman, B. S. (1994). Meta-analysis and the Solomon four-group design. *Journal of Experimental Education, 62*(4), 361-376. doi: 10.1080/00220973.1994.9944140

Sawilowsky, S. S., & Markman, B. (1988). Another look at the power of meta-analysis in the Solomon four-group design. Retrieved from ERIC database. (ED316556)

Sawilowsky, S. S., & Markman, B. S. (1990a). Another look at the power of meta-analysis in the Solomon four-group design. *Perceptual and Motor Skills, 71*(1), 177-178. doi: 10.2466/pms.1990.71.1.177

Sawilowsky, S. S., & Markman, B. S. (1990b). Rejoinder to Braver and Walton Braver. *Perceptual and Motor Skills, 71*(2), 424-426. doi: 10.2466/pms.1990.71.2.424

Statistical Analysis Systems Institute, Inc. (1990). *SAS/STAT user's guide* (Vol. 1) (4th ed.). Cary, NC: Statistical Analysis Systems Institute, Inc.

Wilcox, R. R. (1996). *Statistics for the social sciences*. London, UK: Academic Press.

Wilkinson, L. (1990). *SYSTAT: The system for statistics*. Evanston, IL: SYSTAT.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York, NY: McGraw-Hill.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 69-85.
doi: 10.22237/jmasm/1493597040

ISSN 1538 − 9472

Limitations in the Systematic Analysis of Structural Equation Model Fit Indices

Sarah A. Rose
Wayne State University
Detroit, MI

Barry Markman
Wayne State University
Detroit, MI

Shlomo Sawilowsky
Wayne State University
Detroit, MI

The purpose of this study was to evaluate the sensitivity of selected fit index statistics in determining model fit in structural equation modeling (SEM). The results indicated a large dependency on correlation magnitude of the input correlation matrix, with mixed results when the correlation magnitudes were low and a primary indication of good model fit. This was due to the default SEM method of Maximum Likelihood that assumes unstandardized correlation values. However, this warning is not well-known, and is only obscurely mentioned in some textbooks. Many SEM computer software programs do not give appropriate error indications that the results are unsubstantiated when standardized correlation values are provided.

Keywords: Structural equation model, SEM, fit indices, RMSEA, SRMR, CFI, covariance matrices

Introduction

Wright (1918) presented the foundational theory of Structural Equation Modeling (SEM) for social and behavioral science research based on a path analysis used to model the bone size of rabbits. The novelty of the methodology was more generally accepted a half century later (Matsueda, 2011), coinciding with increasing use of computers, allowing for the more practical use of complicated matrix models. The development of more complicated analytical procedures was inevitable. Hoyle (1995) indicated, "with the increasing complexity and specificity of research questions in the social and behavioral sciences…has come increasing interest in SEM as a standard approach to testing research hypotheses" (p. 1).

Dr. Rose is an Adjunct Instructor of Educational Evaluation and Research. Email her at: ak1734@wayne.edu. Dr. Markman is a Professor of Educational Psychology and Educational Evaluation and Research. Email him at: barry.markman@wayne.edu. Dr. Sawilowsky is a Professor of Educational Evaluation and Research. Email him at: professorshlomo@gmail.com.

SEM is a powerful set of tools that can be used to explore data for the purpose of improving the understanding of the social, psychological, educational constructs and their interactions. It allows for a more complete and comprehensive analysis compared to other research methodologies, because it allows freedom in the evaluation of several model construct relationships simultaneously (Alavifar, Karimimalayer, & Anuar, 2012). The promise of this advantage should not be underestimated. The ability to take many variables and analyze them together using one test without the necessity for Bonferonni or similar corrections allows for considerable flexibility.

SEM models are developed by determining relationships between observed and/or latent variables to specify an initial model. The model is first analyzed to determine whether it is an appropriate approximation of the data construct. If the model is concluded to be an appropriate approximation, it is further analyzed to ascertain the magnitude and direction of relationships between the different variables.

As SEM was developed, it was designed primarily for the use of analysis of social and behavioral science data. Hence, the boundary conditions for performing SEM and determining model fit are steeped in the conditions typical of social and behavioral sciences, which includes multivariate normality (Gullen, 2000; Kline, 2011; Reinartz, Echambadi, & Chin, 2002; Tomarken & Waller, 2005). However, due to the capability of improving quality of life by analyzing data for complex research studies, SEM is increasingly being used in physical science research (e.g. Kelly, 2011; Ewing, Hamidi, Gallivan, Nelson, & Grace, 2014).

Problem Statement

The purpose of this study is to evaluate the sensitivity of selected fit index statistics in determining model fit. There are similarities between social and behavioral science and physical science data that make this transfer of methodologies apparently appropriate. Both data sets are parametric, can be assigned descriptive statistic values, can be formulated to provide frequency diagrams, and can be used with nonparametric tests. However, physical science data differ from the social behavioral science in several ways. In particular, physical science data typically have different distributions than that of social and behavioral science (e.g., Bradley, 1977, 1982; Ito, 1980; Micceri, 1989; Sawilowsky, Blair, & Micceri, 1990; Tan, 1982). Hence, the question arises: how well would SEM perform using non-normally distributed data commonly found in physical science data? However, an important preliminary step, the purpose of

this study, is to compare how various SEM fit indices work under standard normal conditions.

Model Fit

As the model is created, or specified, a foundational aspect of the SEM is to determine how well the model specified represents the data. It is imperative to specify the best model for the data to gain meaningful results. Model fit indices were developed to quantitatively and objectively assess the model fit. The matter of how to develop the fit statistics and which are the best to use has been a topic of great discussion. Kline (2011) indicated, "For at least 30 years the literature has carried an ongoing discussion about the best ways to test hypotheses and assess model fit" (p. 190).

There are dozens of fit indices measuring fit in a variety of ways. The plethora of indices presents two advantages: (1) They are useful for determining the performance of the model. SEM that is an improper fit to the data would provide inaccurate or erroneous results. (2) The complexity of variable matrices and sheer volume of analysis required point to a necessity for numerous fit index models. As the process is rigorous and complicated, so too the fit indices are difficult to simplify. It is therefore not surprising that currently no single fit index encompasses all the different indices in one comprehensive test (Gullen, 2000).

The complexity of analyzing the fit indices and the plethora of index tests from which to form a model fit assumption make it necessary to determine when models are truly a good fit to the data. Hooper, Coughlan, and Mullen (2008) indicated:

> Given the plethora of fit indices, it becomes a temptation to choose those fit indices that indicate the best fit…This should be avoided at all costs as it is essentially sweeping important information under the carpet. (p. 56)

Model fit indices have a short but rabid history. Initially, Chi-squared tests were used; however, the test was proved ineffectual due to the large sample sizes that are required for SEM analysis (Gullen, 2000). The Chi-squared test can be comparatively grossly underpowered for tiny data sets and fail to reach statistical significance. It can also be comparatively super-powered for huge data sets, reaching statistical significance in the presence of negligible differences (see, e.g., Kline, 2011, p. 201).

Various alternatives were therefore developed to supplement the model fit analysis (Bollen, 1989). Fit indices are classified into two categories: (1) Model Test Statistic, and (2) Approximate Fit Index (Kline, 2011).

Model Test Statistics and Chi-Squared

In the model test statistic, data are compared with a baseline model which is a covariance matrix of a sample from the data. If the covariance matrix of the overall data matches the covariance matrix of the sample, the model is considered a good fit. If the matrices differ, the discrepancies using the model need to be explained (Kline, 2011).

Model test statistics are typically developed as a "badness-of-fit" (Kline, 2011, p. 193) test. This means that failure to reject the null hypothesis indicates a good fit. Therefore, it is preferable for the resultant model test statistic to be as small as possible. The basic model test statistic is the model Chi-squared test. This test was developed by Karl Pearson (1900) and has withstood the test of time. It is probably the most well-known and accepted fit statistic. Its value lies in that it is nonparametric. The formula is (Neave & Worthington, 1988):

$$\chi^2 = \sum \frac{\left(\text{Observed} - \text{Expected}\right)^2}{\text{Expected}} \tag{1}$$

Therefore, the Chi-squared statistic is a percentage of the squared deviation from the expected over the expected score. A large Chi-squared statistic indicates a large deviation from the expected distribution. Indication of poor model fit occurs when the Chi-squared statistic value is greater than the critical value based on the nominal alpha.

Although the Chi-squared statistic in this context is apparently nonparametric, there are several factors that can adversely impact it such as large correlations among variables, unique variance, and large sample size (Kline, 2011). When observed variables are highly correlated, the Chi-squared value tends to increase. Unique variances among variables, being a product of score unreliability, result in a loss of statistical power. As the Chi-squared test is a badness-of-fit test, the loss of power reduces the probability of determining a poor model fit. As indicated above, the Chi-squared value tends to increase with sample size.

Approximate Fit Indices

The second type of fit statistic is the approximate fit index. The difference between approximate fit indices and model test statistics is that fit statistics are based on continuous measures. There is not a dichotomous conclusion to either reject or accept a null hypothesis. The value of the fit statistic, as it compares to an ideal value in magnitude, provides a representation of the fit. For example, the ideal value for CFI fit index is 1.0. A model resulting in a CFI of 0.90 would be a better fit than a model resulting in a CFI value of 0.85. As the null hypothesis is not rejected at a decided alpha value, the magnitude of the value has meaning. Therefore, these fit indices can be considered as "rules-of-thumb" as opposed to "golden rules" (Kline, 2011, p. 197).

Approximate fit indices do not "distinguish between what may be sampling error and what may be real covariance evidence against the model" (Kline, 2011, p. 195). Thus, they do not provide information in regards to specification error. These tests are typically goodness-of-fit tests, which mean the ideal index statistic occurs at a value of a specified magnitude (e.g., 1.0 as opposed to zero). The most common of the approximate fit indices are RMSEA, SRMR and the CFI.

Root Mean Square Error Approximation (RMSEA)

The RMSEA is a parsimony-adjusted index. It is not a measure of central tendency but follows a non-central Chi-squared distribution. It has a high and a low value that are provided by most SEM software. The RMSEA is a badness-of-fit test. Therefore, a good fit indicator occurs when the RMSEA low value is less than 0.05 and the high value is less than 0.10. (Kline, 2011).

As a parsimony-adjusted index, the RMSEA adjusts for parsimonious characteristics. It is obtained by dividing by degrees of freedom of the SEM model (Kline, 2011):

$$\text{RMSEA} = \sqrt{\frac{\chi^2_{\text{M}} - df_{\text{M}}}{df_{\text{M}}\left(N-1\right)}} \tag{2}$$

where df_{M} = degrees of freedom of the SEM, N = sample size, and χ^2_{M} = Chi-squared statistic value.

A small Chi-squared value indicates a good model fit. A model with a large degree of freedom, or a parsimonious model, results in a small RMSEA value. In other words, parsimonious models that have small deviations would indicate a

good model fit per this index. The equation is further divided by the sample size. Therefore, the parsimonious effect of the equation increases as sample size increases.

The limitations of RMSEA are obvious. It contains inherent prejudices towards models that have large sample sizes and large degrees of freedom. A model with a moderate-to-large variation from the expected values, but with a large sample size, could pass the RMSEA criteria for model fit.

Standardized Root Mean Square Residual (SRMR)

Although the name is similar to the RMSEA, the two indices are quite different (Iacobucci, 2009). The SRMR is a measure of the standardized value of the square root of the mean absolute covariance squared residual. A good fit value would be close to zero. Hu and Bentler (1999) opined a maximum allowable value for a good fit is approximately 0.09.

The formula, as given by Iacobucci (2009) and Schermelleh-Engel, Moosbrugger, & Muller (2003), is

$$\text{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left[\frac{\left(s_{ij}-\hat{\sigma}_{ij}\right)}{\left(s_{ii}s_{jj}\right)}\right]^2}{\frac{k\left(k+1\right)}{2}}} \tag{3}$$

where k = observed endogenous variables + observed exogenous variables, s_{ij}, s_{ii}, and s_{jj} = values from the covariance matrix, and $\hat{\sigma}_{ij}$ = value from the expected matrix covariance.

Comparative Fit Index (CFI)

The CFI is an incremental fit index and a parsimony-adjusted index, where the data set is compared to the Chi-squared values of a baseline model. It performs well even with small sample sizes. It is a goodness-of-fit test where a value of 1 indicates the best fit. The CFI was developed with the assumption that latent variables are not correlated (Hooper, Coughlan, & Mullen, 2008). Therefore, models with highly correlated latent variables can result in an inaccurate assessment of model fit.

The CFI is a function of the Chi-squared value and degrees of freedom of the model. The formula is (Kline, 2011):

$$\mathrm{CFI} = 1 - \frac{\chi^2_{\mathrm{M}} - df_{\mathrm{M}}}{\chi^2_{\mathrm{B}} - df_{\mathrm{B}}} \tag{4}$$

where df_{X} = degrees of freedom of the SEM/Baseline models, χ^2_{X} = Chi-squared statistic value for the SEM/Baseline models, M = SEM model, and B = baseline model. This equation results in higher values for models with larger degrees of freedom, resulting in a more favorable fit statistic. Hu and Bentler (1999) opined a minimum CFI of 0.95 is necessary to indicate an acceptable fit.

Model Fit Indices Overview

Although multivariate normality is a baseline assumption of the model fit indices (Kline, 2011; Schermelleh-Engel et al., 2003), the formulas for calculating the model fit statistics are apparently nonparametric. It would therefore be reasonable to assume that the model fit index equations could be used to assess model fit for any distribution. However, the robustness of the formulas have not yet been assessed, and the capability of the indices to measure model fit for physical science data is of great interest.

Methodology

Monte Carlo simulation theory requires that baseline theories be tested prior to performance of Monte Carlo simulations on the problem statement. Therefore, it is required to verify model fit indices when normality is not violated as a prerequisite to any study on models that violate underlying assumptions.

Monte Carlo simulations using correlation matrices of randomly selected values of an incrementally increasing correlation range was conducted. The correlation matrices were of randomly selected values, of no model, and no relationship. Model fit indices should indicate a poor model fit for all simulations, meaning they should not exceed the Type-I error rate dictated by nominal α. Therefore, assessment of legitimacy of the model fit index results was based on the percentage of times the results indicated a poor model fit.

At first a Monte Carlo was performed using RStudio based on four variables and 10,000 repetitions of varying correlation matrices of randomly selected numbers between negative and positive 0.1. The results from this simulation

series were mixed in terms of model fit, indicating meaningless results. It was therefore a matter of interest to determine the minimum allowable correlation values under which the model fit indices would provide legitimate and meaningful results.

Monte Carlo simulations based on four variables and correlation matrices of randomly selected values of incrementally-increasing correlation ranges were performed. Each Monte Carlo simulation contained 1,000 repetitions and was performed for sample size of n = 50, 100, 150, 200, 300, and 500. The correlation range was a base value $\pm$ 0.015. Base values were incrementally increased at every hundredths place, beginning from 0.04, 0.05,…, 0.26, and 0.27. As such, 24 Monte Carlo Simulations were performed for six different sample sizes.

Results

Minimum Correlation Coefficient for SEM

The first Monte Carlo simulation included a correlation matrix of random values from a range of 0.04 ± 0.015. All model fit indices results included in the analyses (Chi-squared, RMSEA Lower, RMSEA Upper, SRMR, and CFI) were an indication of a poor model fit 0% of the time. Refer to Table 1 below.

As the correlation matrix values were increased in magnitude, the results of the model fit indices became meaningless. The percentages of greater than and less than critical values did not result in percentage numbers that added to 100%. The fit index results ceased to be meaningless as the correlation magnitudes were continuously increased, and instead were an indication of a poor model fit with increasing reliability. At a certain correlation magnitude (e.g. when correlation was equal to 0.08 ± 0.015 as in Table 2), the results of the model fit indices were an indication of a poor model fit for the conditions studied for all Monte Carlo repetitions. A summary of these results (select simulations) is provided in Table 3.

Each model fit index resulted in legitimate results at different correlation magnitudes. The best model fit index, which resulted in legitimate model fit estimation at the lowest correlation magnitude, was RMSEA Upper at a correlation of 0.08 for all sample sizes. The next best model fit index was CFI, with valid estimation of model fit at a minimum correlation value of 0.16. The next best model fit index was SRMR, with valid model fit estimation at a minimum correlation value of 0.17 for large sample sizes and 0.18 for sample size of 50. The next best model fit index following SRMR was Chi-squared, with valid model fit estimation at a minimum correlation value of 0.24. The model fit index

that performed the poorest was RMSEA Lower, with valid model fit estimation at a minimum correlation of 0.27. Refer to Table 4 below.

Table 1. Monte Carlo simulation percentage of model fit indices (indication of poor model fit); correlation matrix magnitudes range of 0.04 ± 0.015

	Sample Size					
Model Fit Index	**50**	**100**	**150**	**200**	**300**	**500**
Chi-squared	0%	0%	0%	0%	0%	0%
RMSEA Lower	0%	0%	0%	0%	0%	0%
RMSEA Upper	0%	0%	0%	0%	0%	0%
SRMR	0%	0%	0%	0%	0%	0%
CFI	0%	0%	0%	0%	0%	0%

Table 2. Monte Carlo simulation percentage of model fit indices (indication of poor model fit); correlation matrix magnitudes range of 0.08 ± 0.015

	Sample Size					
Model Fit Index	**50**	**100**	**150**	**200**	**300**	**500**
Chi-squared	0%	0%	0%	0%	N/A	N/A
RMSEA Lower	0%	0%	0%	0%	0%	0%
RMSEA Upper	100%	100%	100%	100%	100%	100%
SRMR	0%	0%	0%	0%	0%	0%
CFI	0%	0%	N/A	N/A	N/A	N/A

Table 3. Monte Carlo simulation percentage of model fit indices (indication of poor model fit); correlation matrix magnitudes range ± 0.015

		Sample Size					
Model Fit Index	**Correlation**	**50**	**100**	**150**	**200**	**300**	**500**
Chi-squared	0.04	0%	0%	0%	0%	0%	0%
RMSEA Lower		0%	0%	0%	0%	0%	0%
RMSEA Upper		0%	0%	0%	0%	0%	0%
SRMR		0%	0%	0%	0%	0%	0%
CFI		0%	0%	0%	0%	0%	0%
Chi-squared	0.06	0%	0%	0%	0%	0%	N/A
RMSEA Lower		0%	0%	0%	0%	0%	0%
RMSEA Upper		35%	N/A	N/A	N/A	N/A	N/A
SRMR		0%	0%	0%	0%	0%	0%
CFI		0%	0%	0%	N/A	N/A	N/A
Chi-squared	0.08	0%	0%	0%	0%	N/A	N/A
RMSEA Lower		0%	0%	0%	0%	0%	0%
RMSEA Upper		100%	100%	100%	100%	100%	100%
SRMR		0%	0%	0%	0%	0%	0%
CFI		0%	0%	N/A	N/A	N/A	N/A

Table 3, continued.

		Sample Size					
Model Fit Index	**Correlation**	**50**	**100**	**150**	**200**	**300**	**500**
Chi-squared	0.16	0%	N/A	N/A	N/A	N/A	N/A
RMSEA Lower		0%	0%	N/A	N/A	N/A	N/A
SRMR		54%	N/A	N/A	N/A	N/A	N/A
CFI		100%	100%	100%	100%	100%	100%
Chi-squared	0.18	0%	N/A	N/A	N/A	N/A	N/A
RMSEA Lower		0%	N/A	N/A	N/A	N/A	N/A
SRMR		100%	100%	100%	100%	100%	100%
Chi-squared	0.24	100%	100%	100%	100%	100%	100%
RMSEA Lower		0%	N/A	N/A	N/A	N/A	N/A
Chi-squared	0.27	100%	100%	100%	100%	100%	100%
RMSEA Lower		100%	100%	100%	100%	100%	100%
RMSEA Upper		100%	100%	100%	100%	100%	100%
SRMR		100%	100%	100%	100%	100%	100%
CFI		100%	100%	100%	100%	100%	100%

Table 4. Minimum correlation values for valid model fit index measurement

	Sample Size					
Model Fit Index	**50**	**100**	**150**	**200**	**300**	**500**
Chi-squared	0.24	0.24	0.24	0.24	0.24	0.24
RMSEA Lower	0.27	0.27	0.27	0.27	0.27	0.27
RMSEA Upper	0.08	0.08	0.08	0.08	0.08	0.08
SRMR	0.18	0.17	0.17	0.17	0.17	0.17
CFI	0.16	0.16	0.16	0.16	0.16	0.16

Conclusion

Originally, a Monte Carlo simulation with randomly selected correlation values between - 0.1 and + 0.1 was performed. The results were meaningless, with mixed results in terms of fit. The output of the latest repetition of the Monte Carlo simulation was extracted and compared with the output from Amos Graphics to ensure that a programming error did not occur. The results were the same within rounding error.

Fit index results should be consistent regardless of whether or not a meaningful model is produced. Examination of the model fit results should indicate a good or a poor model fit when a reasonable model is assessed. However, examination of the results should never indicate a good model fit on a poorly-defined model. In this case, the correlation values between variables were small and the paths were not significant. Therefore the model, having no relationships,

should result in an indication of poor model fit when assessed using model fit index tests. This indication of poor model fit should occur uniformly for all model fit index tests and for all sample sizes, or at least within the Type I error rate set by nominal alpha.

These findings were discussed with colleagues. One believed that, with caution (presumably ignoring fit results in the absence of a good model), there were some insights that could be garnered based on the results. This viewpoint was amplified by another colleague, who replicated the results via Mplus, and hence urged extreme caution, because of SEMs ability to produce a well-fitted model that is nevertheless bereft of significant covariances.

As a beginning to approaching the model fit assessment with caution, additional research was conducted to determine what SEM conditions caused the model fit index results to be meaningless. The Monte Carlo simulation models were assessed to discover common characteristics. A consistent attribute was the low correlation values between the variables. It appeared when the correlation values between variables were low, the results of the model fit indices were meaningless. Additional research was therefore conducted to determine what constituted a low correlation, and whether there was a minimum allowable correlation value between variables that is a prerequisite for a SEM to be meaningful.

Additional Monte Carlo simulations were conducted, with 1,000 repetitions and varying magnitudes of correlation matrices. The magnitudes of the correlation values were randomly selected from a base value $\pm$ 0.015. Twenty-four Monte Carlo simulations were performed, with the base value increasing from 0.04 to 0.27 at every hundredths place value (i.e. 0.04, 0.05 ,0.06, etc.). The model fit indices would be legitimized by the percentage of times a poor model fit was indicated, as the variables had no relationship and correlation values were randomly selected.

As the correlation matrix values were increased in magnitude, the results of the model fit indices became first illogical and then finally logical with an indication of a poor model fit occurring with increasing reliability. At a certain correlation magnitude range (e.g. when correlation was equal to 0.08 $\pm$ 0.015 as in Table 2), the results of the model fit indices were an indication of a poor model fit for all sample sizes studied for all Monte Carlo repetitions.

Table 5. Minimum correlation values

Rank	Model Fit Index	Minimum Correlation Value
1	RMSEA Upper	0.08
2	CFI	0.16
3	SRMR	0.18
4	Chi-squared	0.24
5	RMSEA Lower	0.27

Table 6. Correlation matrix

Variables	*z*	X_1	X_2	X_3
z	1.000	0.104	0.098	0.115
X_1	0.104	1.000	0.100	0.088
X_2	0.098	0.100	1.000	0.109
X_3	0.115	0.088	0.109	1.000

Each model fit index resulted in legitimate results at different correlation magnitudes; refer to Table 3 above. Model fit indices can be ranked from best to worst based on the minimum correlation values required before legitimate results were acquired. The model fit indices, from best to worst, are listed in Table 5 above with their respective minimum correlation values and ranks.

The results from the last repetition of the Monte Carlo simulation with correlation range of 0.1 ± 0.015 and sample size of 500 were extracted (refer to Table 6 above and the Lavaan output below) to better understand the results of the Monte Carlo simulations and to verify the conclusions determined above. The results of the model fit index tests were mixed. The *p*-value for the Chi-squared test was 0.003, an indication of a poor model fit. The RMSEA Upper value was 0.133, an indication of a poor model fit. The RMSEA Lower value was 0.044, an indication of a good model fit. The CFI value was 0.505, an indication of a poor model fit. The SRMR value was 0.055, an indication of a good model fit.

The regression coefficients for the exogenous variables were 0.088 for X_1, 0.079 for X_2, and 0.098 for X_3. Although these values were low, the coefficients for X_1 and X_3 were statistically significant. This is illogical, as the correlation magnitudes in the correlation matrix were all low. Statistically significant paths between variables are therefore a contradictory conclusion. These results solidified the conclusion above that a SEM with a correlation matrix of low values would result in illogical outcomes.

Lavaan Output for Sample Size of 500 and Four Variables, Repetitions = 1,000

```
Number of observations           500
Estimator                        ML
Minimum Function Test Statistic  14.059
Degrees of freedom               3
P-value (Chi-square)             0.003

User model versus baseline model:
 Comparative Fit Index (CFI)     0.505
 Tucker-Lewis Index (TLI)        0.010
 Number of free parameters       7
 RMSEA                           0.086
rmsea.ci.lower                   0.044
rmsea.ci.upper                   0.133
 90 Percent Confidence Interval  0.044 0.133
 P-value RMSEA <= 0.05           0.075
 SRMR                            0.055

Parameter estimates:
 Information        Expected
 Standard Errors    Standard

Regressions:
z ~    Estimate     Std.err      Z-value      P(>|z|)
x1     0.088        0.044        1.990        0.047
x2     0.079        0.044        1.786        0.074
x3     0.098        0.044        2.232        0.026

Covariances:
 x1 ~~x2     0.000
 x3          0.000
 x2 ~~x3     0.000

Variances:
 z     0.970  0.061
 x1    0.998  0.063
 x2    0.998  0.063
 x3    0.998  0.063
```

SEM is a collection of procedures that are assessed based on a plethora of fit or lack of fit statistics that could be subjectively chosen or ignored to support or eliminate a model. Dozens of caveats (such as those listed in Kline, 2011, e.g., at its core it relates to non-experimental data and hence there can never be causation (p. 8), a poor model can be preserved by modifying the hypotheses on which it is based (p. 8), alternative models may not be ruled out (p. 8), it is a large sample technique (p. 11), it eschews hypothesis testing and hence is veiled behind subjectivity (p. 13), the statistical significance of estimated parameters are dependent on the algorithm adopted (p. 13), a maximum likelihood estimate cannot tolerate even a single missing datum (p. 48), a nonpositive definite matrix cannot be analyzed (p. 49), ill-scaled covariance matrices cannot be handled (p. 67)) severely limit SEM outside of textbook examples.

Moreover, Kline (2011) noted,

> It may be problematic to submit for analysis just a correlation matrix without standard deviations or specify that all standard deviations are 1.0, which standardizes everything. This is because the default method of ML estimation (and most other methods, too) assumes that the variables are unstandardized. This means that if a correlation matrix without standard deviations is analyzed, the results may not be correct…Some SEM computer programs give warning message or terminate the run if the researcher requests the analysis of a correlation matrix only with standard ML estimation. By the same token, it would also be problematic to convert raw scores to *z* scores and then submit for analysis the data file of standardized scores. (p. 49)

These cautions from Kline (2011) appear to explain why a systematic Monte Carlo study conducted by inputting an incrementally increasing correlation matrices, such as was attempted in this study, cannot be successful. The standard procedure of starting the study with a null zero order correlation matrix to show the relevant fit indices reject, or fail to reject as appropriate to the index, is not possible, precluding a presentation of the power spectrum of the competitors based on systematically increasing (or decreasing based on the type of fit index) the matrix. The restrictions indicated by Kline (2011) were mentioned in an obscure section of the textbook, and were omitted by most other textbook authors. Thus, this limitation and the egregious results from the non-compliance are not well-publicized.

It appears it is necessary to start with a good model in order for the model fit indices to provide a proper assessment. This is circuitous, for how can a good model be assessed if the baseline condition for meaningful results is a good model? Analysts must consider this paradox, and decide if SEM outside of textbook examples is truly meaningful.

References

Alavifar, A., Karimimalayer, M., & Anuar, M. K. (2012). Structural equation modeling VS multiple regression. *Engineering Science and Technology: An International Journal, 2*(2), 326-329. Retrieved from http://www.estij.org/papers/vol2no22012/25vol2no2.pdf

Bollen, K. A. (1989). A new incremental fit Index for general structural equation models. *Sociological Methods & Research, 17*(3), 303-316. doi: 10.1177/0049124189017003004

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician, 31*(4), 147-150. doi: 10.2307/2683535

Bradley, J. V. (1982). The insidious L-shaped distribution. *Bulletin of the Psychometric Society, 20*(2), 85-88. doi: 10.3758/BF03330089

Ewing, R., Hamidi, S., Gallivan, F., Nelson, A. C., & Grace, J. B. (2014). Structural equation models of VMT growth in US urbanised areas. *Urban Studies, 51*(14), 3079-3096. doi: 10.1177/0042098013516521

Gullen, J. A. (2000). *Goodness of fit as a single factor structural equation model* (Unpublished doctoral dissertation). Wayne State University, Detroit, MI.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60. Retrieved from http://www.ejbrm.com/volume6/issue1

Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Iacobucci, D. (2009). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*(1), 90-98. doi: 10.1016/j.jcps.2009.09.003

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of Statistics* (Vol. 1, pp. 199-236). Amsterdam, Netherlands: North-Holland. doi: 10.1016/S0169-7161(80)01009-7

Kelly, S. (2011). Do homes that are more energy efficient consume less energy?: A structural equation model of the English residential sector. *Energy, 36*(9), 5610-5620. doi: 10.1016/j.energy.2011.07.009

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Matsueda, R. L. (2011). *Key advances in the history of structural equation modeling* (Working paper no. 114). Seattle, WA: University of Washington Center for Statistics and the Social Sciences. Retrieved from https://www.csss.washington.edu/Papers/2012/wp114.pdf

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

Neave, H. R., & Worthington, P. L. (1988). *Distribution free tests*. Boston, MA: Unwin Hyman Inc.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5, 50*(302), 157-175. doi: 10.1080/14786440009463897

Reinartz, W. J., Echambadi, R., & Chin, W. W. (2002). Generating non-normal data for simulation of structural equation models using Mattson's method. *Multivariate Behavioral Research, 37*(2), 227-244. doi: 10.1207/S15327906MBR3702_03

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). REALPOPS.LIB: a PC Fortran library of eight real distributions in psychology and education. *Psychometrika, 55*(4), 729.

Schermelleh-Engel, K., Moosbrugger, H., & Muller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8*(2), 23-74. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpr-online/issue20/art2/mpr130_13.pdf

Tan, W. Y. (1982). Sampling distributions and robustness of t, F, and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics – Theory and Methods, 11*(21), 2485-2511.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1*(1), 31-65. doi: 10.1146/annurev.clinpsy.1.102803.144239

Wright, S. (1918). On the nature of size factors. *Genetics, 3*(4), 367-374. Retrieved from http://www.genetics.org/content/3/4/367

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 86-106.
doi: 10.22237/jmasm/1493597100

ISSN 1538 − 9472

Guidelines for Generating Right-Censored Outcomes from a Cox Model Extended to Accommodate Time-Varying Covariates

Maria E. Montez-Rath
Stanford University
Stanford, CA

Kristopher Kapphahn
Stanford University
Stanford, CA

Maya B. Mathur
Stanford University
Stanford, CA

Aya A. Mitani
Boston University
Boston, MA

David J. Hendry
London School of Economics
London, UK

Manisha Desai
Stanford University
Stanford, CA

Simulating studies with right-censored outcomes as functions of time-varying covariates is discussed. Guidelines on the use of an algorithm developed by Zhou and implemented by Hendry are provided. Through simulation studies, the sensitivity of the method to user inputs is considered.

Keywords: Right-censored outcomes, extended Cox model, time-varying covariates, simulation studies, censoring distribution

Introduction

The development and evaluation of methods for data analysis are often facilitated through simulation studies, particularly when closed-form solutions are unknown (Burton, Altman, Royston, & Holder, 2006). Simulation studies can be especially useful for assessing the behavior of analytic techniques under various conditions that present complexities in practice. For example, Collins, Schafer, and Kam (2001) described the bias that resulted from multiple imputation methods that utilized varying degrees of auxiliary data by simulating data under conditions that varied the percentage missing, the reasons for missingness, and the strength and availability of auxiliary information. Desai, Bryson, and Robinson (2013) performed a simulation study to evaluate properties of robustly-estimated standard errors in the presence of clustering when clustering membership is mis-specified. In research to evaluate and develop methods for handling missing data,

Dr. Montez-Rath is a Biostatistician in the Division of Nephrology. Email her at: mmrath@stanford.edu.

simulating studies with right-censored outcomes as functions of time-varying covariates is critical. There is particular interest in simulating studies with characteristics, including correlation structures over time and across features, that closely resemble a complicated motivating data set.

A large body of research has been devoted to generating right-censored survival times from time-invariant covariates. For example, Leemis, Shih, and Reynertson (1990) demonstrated that survival times that followed a Cox proportional hazards model could be generated by inverting the cumulative hazard function. Independently, Bender, Augustin, and Blettner (2005) offered details on simulating survival times from such a model where the hazard function was assumed to follow exponential, Weibull or Gompertz distributions.

However, generating right-censored outcomes as functions of time-varying covariates is more complicated; a subject's outcome corresponds to multiple values of a covariate over time where the number of values for the covariate may vary across subjects. Using the approach described by Bender et al. (2005) for this purpose is challenging as it would require inversion of the expression $-H_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}(t))$ which includes the cumulative hazard function. Sylvestre and Abrahamowicz (2008) argue that such inversion cannot be easily done since it is only possible if the baseline hazard can be represented by a parametric function. A possible solution is to express changes over time in the covariate, $\mathbf{x}(t)$, as a parametric function that is well-defined over the range of time studied. To that end, Austin (2012) extended the work of Bender et al. (2005) although the extension is limited in that it can only accommodate one time-varying covariate.

Alternatively, Sylvestre and Abrahamowicz (2008) evaluated extending an algorithm first introduced by Abrahamowicz, MacKenzie, and Esdaile (1996) for time-invariant covariates; this algorithm did not require inverting the cumulative hazard function. Instead, the algorithm matches, one-to-one, survival times and covariates that have been generated independently, based on a probability law derived from the partial likelihood of the Cox proportional hazards model. This method allows for any number of time-invariant as well as time-varying covariates without a need to specify a functional form for how they vary over time, but the proposed process of generating the survival times has no closed-form solution. Time-dependent effects, i.e., effects that would vary depending on the time interval, can be introduced directly in the vector of survival times provided to the algorithm but generating those survival times is challenging. Similarly, Crowther and Lambert (2013) proposed a method that relies on numerical integration and allows explicit modeling of the baseline and estimation of the

absolute hazard but it can be computationally expensive if the number of covariates is large.

Independently, Zhou (2001) showed that right-censored outcomes can be generated by transforming a random variable that follows a piecewise exponential distribution, where the hazard is assumed to be constant within a time interval but can vary across time intervals that are defined by changes in the covariate. A closed-form solution for generating the data was also provided. Hendry (2014) developed a general algorithm (with code in R) that implements Zhou's method to generate right-censored survival times under the Cox model with any number of both time-invariant and time-varying covariates that vary at integer-valued steps of the time scale.

This study focuses on Zhou's method for three important reasons. The first is that it is supported by readily accessible software developed by Hendry (2014), providing easy access to a wide audience of potential users. The second is that it can accommodate any number of time-invariant and/or time-varying covariates. Finally, although not highlighted in our study here, Zhou's method provides the additional flexibility of enabling relaxation of the proportionality assumption by allowing the effects to vary between time-intervals (time-dependent effects). Note that the latter is not a feature shared by other methods.

There are multiple user-supplied parameters involved in applying Hendry's implementation of Zhou's method, but properties of the distribution of the outcome may be sensitive to their specification. The primary purpose of this paper is to evaluate these sensitivities and provide guidelines on the use of the Hendry algorithm. To that end, based on an extensive simulation study, we suggest a flexible form for the baseline hazard and characterize the sensitivity of the method to other user inputs under a variety of conditions. Specifically, sensitivities of the algorithm to the censoring distribution are addressed, the shape of the hazard, the degree of correlation between covariates, and the type of covariates. The performance of the algorithm is evaluated through standardized bias and mean squared error of the fitted coefficients and use these statistics to inform guidelines on use of the algorithm.

Cox Regression Models with Time-Varying Covariates via the Piecewise-Exponential Distribution

Zhou (2001) showed that if Y_j, $j = 1,\ldots, J$ are random variables that follow a piecewise exponential distribution, where J indicates the number of intervals, and g(.) is a monotone increasing function such that $g(0) = 0$ and $g^{-1}(t)$ is

differentiable, then $g(Y_j)$ follows a Cox model with a time-varying covariate and a baseline hazard $h_0(t) = d/dt[g^{-1}(t)]$. To incorporate covariates, one can specify the piecewise exponential variables with varying rates γ_j such that they depend on any number of time-invariant and/or time-varying covariates $\mathbf{Z}_j = Z_{j1},\ldots, Z_{jP}$ and regression parameters $\boldsymbol{\beta} = \beta_1,\ldots, \beta_P$ where $\boldsymbol{\gamma}_j = \exp(\boldsymbol{\beta}\mathbf{Z}'_j)$. In this form, effect sizes can easily be introduced as the components of the rates for the piecewise exponential variates where the hazard of $g(Y_j)$ is defined by $h_0(t)\exp(\boldsymbol{\beta}\mathbf{Z}'_j)$. Time-dependent effects can be introduced by allowing the effects to vary between time intervals ($\boldsymbol{\beta}_j = \beta_{j1},\ldots, \beta_{jP}$) and so $\gamma_j = \exp(\boldsymbol{\beta}_j\mathbf{Z}'_j)$.

Hendry (2014) demonstrated that piecewise exponential random variables with support $[a, b]$ such that $0 < a < b$ (truncated piecewise exponential random variables), can be generated through an accept-reject algorithm where realizations outside of the support are discarded and those within are included. A full proof for how one can generate survival times that follow a Cox model with time-dependent covariates using a truncated piecewise exponential distribution can be found in Hendry (2014). Key parameters of the algorithm that need to be defined are: the bounds of truncation (a, b), the parameters corresponding to the piecewise exponential random variables or rates γ_j, the transformation function g, and the censoring mechanism.

The bounds of truncation relate to the limits of observed survival times, which can be informed by an empirical data set. For example, a lower bound $a > 0$ can correspond to a lower bound on subject eligibility (e.g., it may be that only subjects who are considered "active" users of a health system – i.e., who exceed a minimum duration of observation – are eligible for study). Note this form of truncation is not to be confused with left truncated time-to-event data, where the latter would constrain observational times to begin at the lower bound. In contrast, here observational times begin at zero but are only included if they exceed the lower bound. The upper bound corresponds to the maximum allowable time observed for an individual. The larger the upper bound, the larger the number of records per individual. This has implications not only for the time needed to generate the data, but also for the run time of any application of the simulated data.

The g function has an important role. It is defined such that $g^{-1}(t) = H_0(t)$, the cumulative baseline hazard of some known function. It should be specified to best represent the disease or process of interest. Options described by Hendry are mostly power functions and tend to lead to large hazards such that events occur soon after the start of observation. Hendry suggests exploring a variety of functional forms to appropriately capture the process studied but does not offer much guidance on parameter choice.

It is recommended the g function be defined through the use of a Weibull distribution so that the variates generated in the process have a baseline hazard of a Weibull random variable. This distribution is a flexible choice defined by two parameters: 1) shape, v, which determines whether the hazard is increasing over time ($v > 1$), constant ($v = 1$), or decreasing ($v < 1$); and 2) scale, λ, which shifts the hazard distribution right or left, depending on the overall survival time. The Weibull distribution has a hazard function defined by $h_0(t) = \lambda v t^{v-1}$ and the cumulative hazard equal to $H_0(t) = \lambda t^v$. By fixing v, one can generate outcomes with a pre-determined median survival time informed by the empirical data (which we will call the target median).

Assume $g^{-1}(t)$ to be the cumulative baseline hazard from the Weibull distribution with shape parameter v and scale parameter λ. The estimated median survival time, $\hat{t}(50)$, for an individual whose vector of explanatory variables is $\mathbf{Z} = (Z_1, \ldots, Z_p)$ with estimated effects $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$, is defined by

$$\hat{t}(50) = \left\{ \frac{\log 2}{\hat{\lambda} e^{\hat{\boldsymbol{\beta}} \mathbf{z}'}} \right\}^{\frac{1}{\hat{v}}}$$

(Collett, 2003, p. 177). This formula can be used to compute a value for λ, given v and a target median. This is done for a hypothetical individual whose covariate values are at the mean $\left(\hat{\boldsymbol{\beta}} \bar{\mathbf{Z}}' = \hat{\beta}_1 \bar{Z}_1 + \ldots + \hat{\beta}_p \bar{Z}_p\right)$.

There are a variety of options that the user can consider for incorporating censoring into the data generation process. Often studies impose administrative censoring where subjects are no longer observed beyond the study end date. This is fairly straightforward to define once times to the event have been generated. However, censoring may arise for other reasons, like when subjects drop out of or withdraw from a study and are lost to follow-up.

There are two main ways to implement this type of censoring. One is referred to as traditional censoring, in which both a survival time and a censoring time are generated and then the minimum value of the two is chosen as the time the subject was observed. If the minimum value was the survival time, an indicator for whether the subject was observed to have the event will equal 1. Otherwise, if the subject's time was censored, the indicator will equal 0. Hendry's algorithm can also be used to impose traditional censoring. To obtain the intended percent of observations being censored, though, additional parameters need to be specified and refined by iteration. The second alternative, referred to as random

censoring, is easier to implement and computationally more efficient. In this approach, each patient's observation is simply censored at random with a probability determined by the percentage of censored observations desired. An indicator for whether the time was censored follows a Bernoulli distribution with a pre-specified probability. For more on incorporating censoring into simulations, see Crowther and Lambert (2013), Burton et al. (2006), and Bakoyannis and Touloumi (2012).

Consider the impact of these parameters on generating data that closely mimic a motivating data set. Specifically, the investigation in this study is on the impact of the parameters for the Weibull distribution, censoring mechanism, correlation among variables, and variable type on properties of estimates obtained from fitting an extended Cox model to data generated using this approach, as well as a generated survival time distributions and variation in computation time.

Methodology

Design of Simulation Study

The parameters of the simulation study follow a full factorial design of the following parameters:

Bounds of truncation (a-b): (20-300), (20-150), (20-50)
Covariate combinations: 2 Normal, $Z_1 \sim N(50, 10^2)$ and $Z_2 \sim N(30, 52)$; 1 Normal + 1 Binary, $Z_1 \sim N(50, 10^2)$ and $Z_2 \sim Bern(0.5)$
Weibull shape parameter (v): 2, 1, 0.5
Target median: 35; 75; 150
Censoring distribution: None; Random; Traditional; Administrative
Percent censored patients (if censoring applied): 20%; 50%; 80%

Data were generated using all possible combinations of the parameters listed with the exception of the percent of patients censored, which is relevant only when an actual censoring distribution is being applied. Details on the choice of parameters are described here.

Data Generated

Using Hendry's algorithm, survival times were generated to fall within 2 bounds of truncation defining the range of possible survival times. The lower bound, *a*, was fixed at 20 and the upper bound, *b*, was allowed to vary (50, 150, and 300).

Survival times depended on two independent time-varying covariates (Z_1, Z_2) in two possible combinations: one in which both covariates are normally distributed random variables and a second in which one covariate is normally distributed and the other is a Bernoulli random variable. Specifically, Z_1 is always assumed to be N(50, 10^2) and Z_2 could either be N(30, 5^2) or Bernoulli(0.5).

The transformation function g was specified to be derived from a Weibull distribution with shape parameter corresponding to an increasing (v = 2), constant (v = 1), or decreasing (v = 0.5) hazard. The scale parameter is computed after providing the shape parameter and the target median survival time, which vary (35, 75, and 150). Note that some target medians fell outside the bounds, demonstrating the impact of parameter choice. For a given shape parameter (v), target median (M), and vector of regression parameters $\boldsymbol{\beta} = (\beta_1, \beta_2)$, the scale parameter λ and the g function are defined as follows:

$$\lambda = \frac{\log 2}{\boldsymbol{\beta}\bar{\mathbf{Z}}'} M^{-v}, \; \mathrm{g}(t) = \left(\lambda^{-1} t\right)^{\frac{1}{v}}, \text{ and } \mathrm{g}^{-1}(t) = \lambda t^{v}$$

where $\bar{\mathbf{Z}} = \left(\bar{Z}_1, \bar{Z}_2\right)$ is the vector of means of the covariates. Under the scenario where $Z_1 \sim \mathrm{N}(50, 10^2)$ and $Z_2 \sim \mathrm{N}(30, 5^2)$ then $\boldsymbol{\beta}\bar{\mathbf{Z}}' = 50\beta_1 + 30\beta_2$. However, if $Z_2 \sim$ Bernoulli(0.5), then $\boldsymbol{\beta}\bar{\mathbf{Z}}' = 50\beta_1 + (0.5)\beta_2$.

The algorithm computes survival times within the defined limits, which might be considered the "true" event times and may or may not be observed depending on the censoring method applied. For administrative and traditional censoring, these are the uncensored times. We then imposed 3 types of censoring (administrative, traditional, and random) with various percentages of patients being censored (20%, 50%, or 80%). In administrative censoring, patients are observed until a fixed time (end of study). In traditional censoring, censoring times were generated in parallel with the uncensored survival times using an independent implementation of the Hendry algorithm. The parameters of the censoring distribution are chosen by iteration to yield the correct amount of censored observations and are different than the parameters used in the creation of the uncensored times, thus reflecting non-informative censoring. The final observed time is defined as the minimum of the two survival times. The event indicator is set to 0 if the censoring time is smaller than the uncensored time. In random censoring, each subject has a probability p_c, set to 0.2, 0.5, or 0.8 (depending on the percentage of censoring desired), of being censored at the end

of the subject's generated survival time. Event indicators were thus distributed as Bernoulli random variables with $p = 1 - p_c$.

The influence of the correlation between the covariates on properties of estimates obtained from fitting the Cox model is examined. The covariates, as defined above, were allowed to be correlated, with correlations ranging from -0.8 to 0.8, using the mvrnorm function in R when generating two Normal random variables and the binnor package in R when generating one Normal and one Bernoulli variable (Demirtas & Doganay, 2012). In this scenario, survival times were set to be bounded between 20 and 300, the shape parameter, v, was fixed at 2, and the target median was fixed at 150.

Number of Replications

For each scenario or combination of the simulation parameters, we drew 1000 simulated data sets (replicates) each with 1000 individuals with varying number of observations per individual depending on the scenario being simulated and the data generated.

Parameters to Be Estimated

We fit the true model (an extended Cox model) to the data generated and obtained estimates for the regression coefficients corresponding to the two covariates. Parameters were set to $\beta_1 = 0.02$ and $\beta_2 = 0.04$ when covariates were two Normal variables and $\beta_1 = 0.02$ and $\beta_2 = -0.5$ when covariates were one Normal and one Bernoulli.

Evaluation Criteria

The performance of the algorithm was assessed by three statistics, which were computed for each parameter estimate $\left(\hat{\beta}_1, \hat{\beta}_2\right)$: the standardized bias (difference between the average estimate and the true value as a percentage of the estimate's empirical standard error), the mean squared error (MSE, squared difference between the true and estimated parameter averaged over the number of simulations), and the coverage percentage (percentage of time the 95% confidence interval contains the true parameter). As suggested by Collins et al. (2001), standardized bias larger than 40% (in absolute value) is considered to indicate poor performance. Although nominal coverage percentage is 95%, Collins and others defined acceptable coverage as 90% or higher. In order to

assess how close the distribution of the generated survival times is to the distribution of times in the empirical data

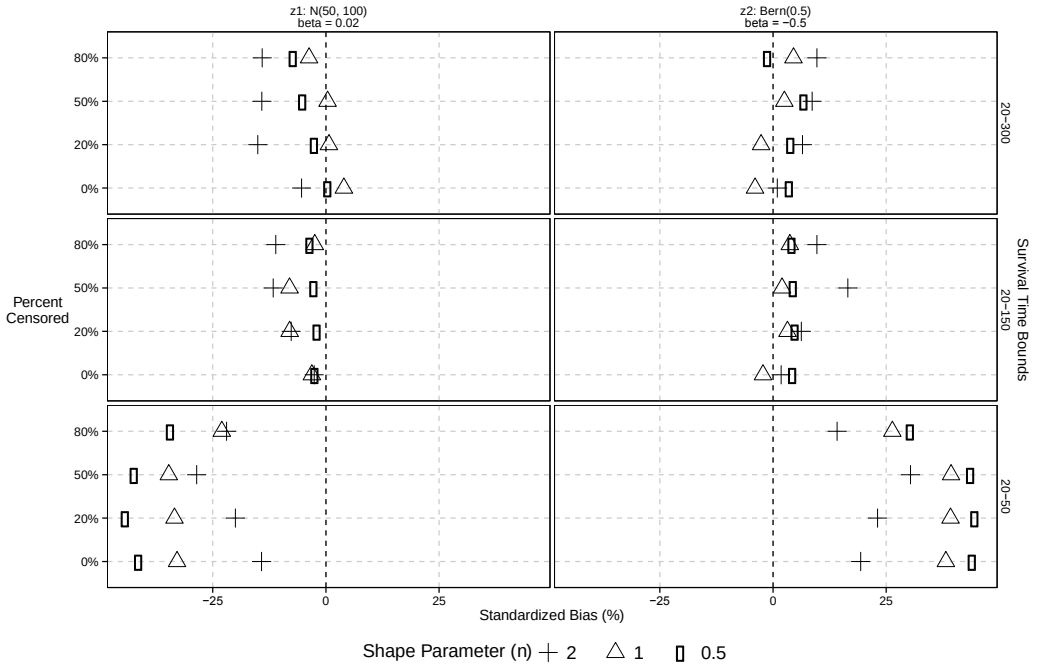


Figure 1. Standardized bias in fitted coefficients when survival times are generated with target median of 35, random censoring, and using mixed covariates

set, we graphically assess the median survival times generated. Finally, we compare algorithm run times across different combinations of simulation parameters.

Results

Impact of Limits on the Survival Times Generated

Data generated under the most restrictive bound (20-50) with a median goal equal to 35, independent of the types of covariates, yielded large standardized bias relative to the other two bounds (e.g., the range of standardized bias for Normal covariates was -48.4 to -14.2, -11.6 to -0.7, and -15.0 to 4.0 for the 20-50, 20-150, and 20-300 bounds, respectively) (Figure 1, left column). For both traditional and administrative censoring, under the most restrictive bound, the standardized bias

decreased as the percent of censored observations increased (e.g., the range of standardized bias for two Normal covariates assuming traditional censoring was -40.9 to -14.2 and -6.3 to -3.2 for 0% and 80% censored, respectively). This was not the case for random censoring, however. For example, the range of standardized bias when the two covariates are normally distributed was -40.9 to -14.2 and -33.9 to -22.0 for 0% and 80% censored observations, respectively (Figure 2, left column). Results are not shown for administrative censoring.

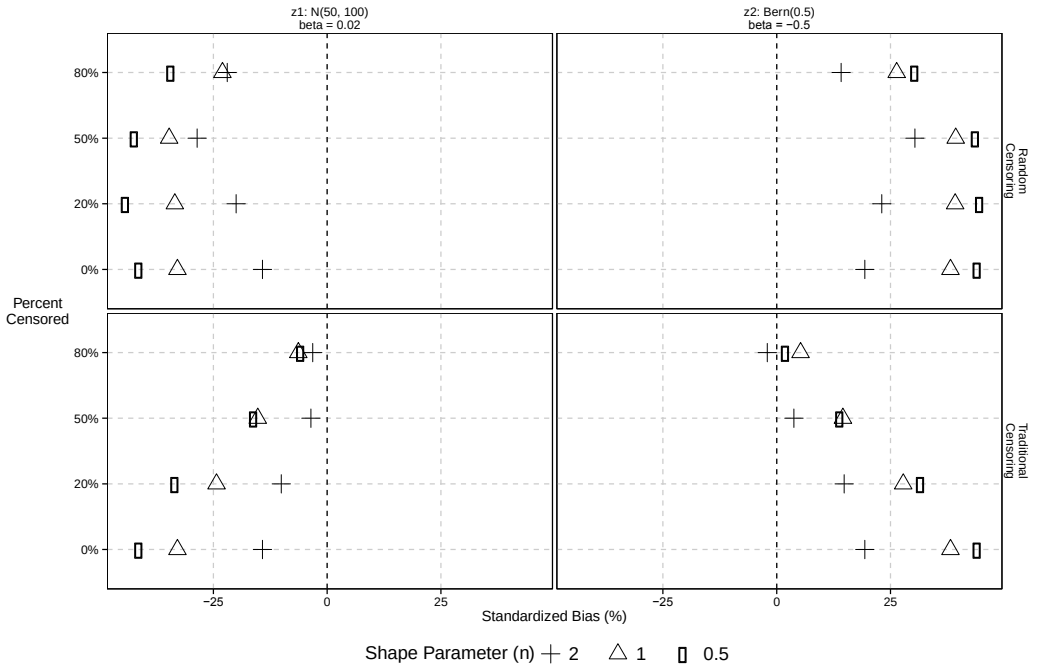


Figure 2. Standardized bias in fitted coefficients when survival times are generated bounded between 20 and 50, target median of 35, and using mixed covariates contrasting random versus traditional censoring

---

Coverage percentages were close to 95% for most combinations of the parameters simulated (0.89-0.97). Somewhat lower coverage – although still over 90% – was obtained when generating times using smaller limits with random censoring (e.g., coverage percentages for 2 Normal covariates assuming a target median of 35 were 0.92 to 0.95, 0.94 to 0.97, and 0.94 to 0.96 for bounds of 20-50, 20-150 and 20-300, respectively) (Figure S1A, middle columns, rows 4-6).

Independent of the bounds and censoring type assumed, when the percent of censored observations increased from 0% to 80% censored for the binary covariate, the MSE increased from 0.004 to 0.027 (Figure S2C, columns 2, 4 and 6, rows 1-3). In contrast, the MSE remained close to zero when covariates followed a Normal distribution (Figure S2C).

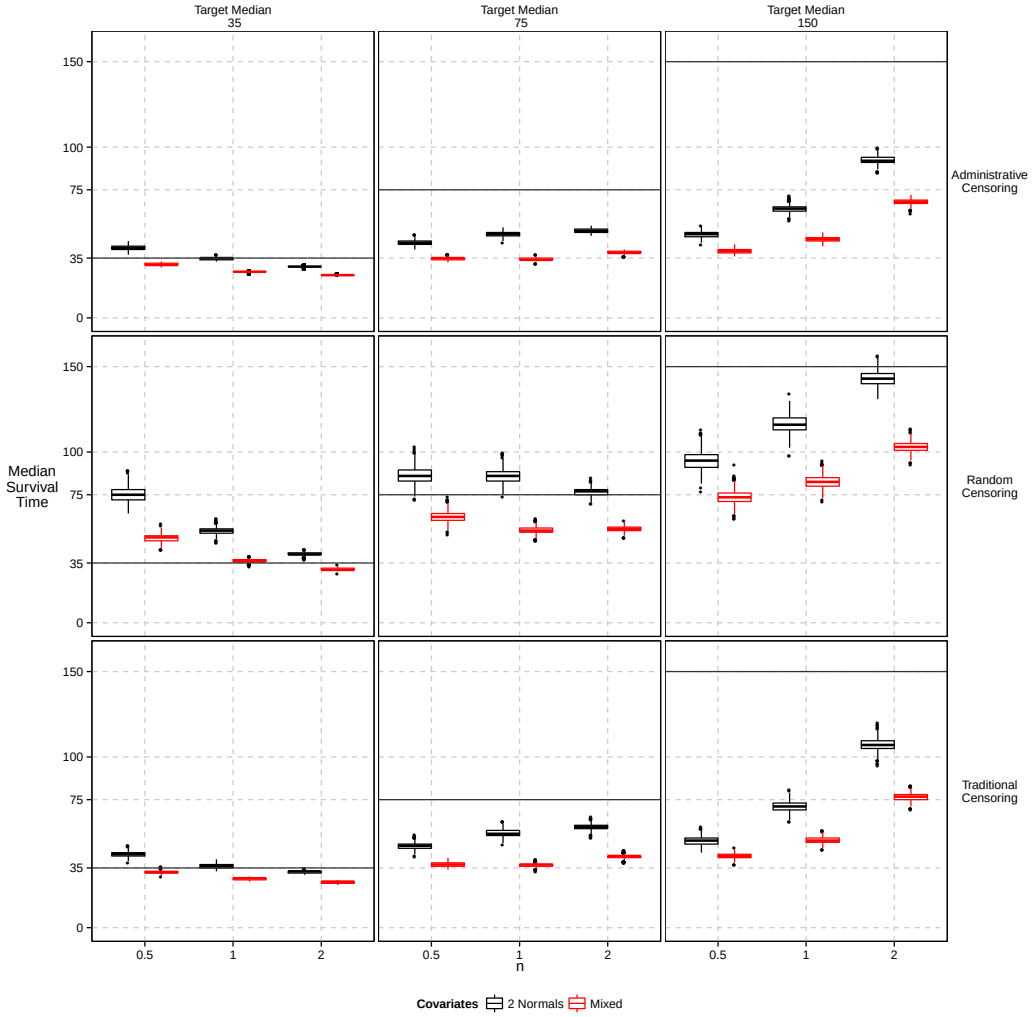


Figure 3. Median survival times for data generated bounded between 20 and 300; straight lines indicate the relevant target median

Impact of g Function Definition

Overall, under data generated with the least restrictive bound (20-300), when the shape parameter for the Weibull distribution was 2 (compared to 1 or 0.5), the median of the generated survival times came closer to the target (Figure 3).

More specifically, under random censoring, a shape parameter of 2 yielded survival time distributions with medians closer to the target median relative to the other shape parameter choices (e.g., for a target median equal to 150, under random censoring with 2 Normal covariates, median survival times ranged from 131.0 to 156.5 and 77 to 113.0 when $v = 2$ and 0.5, respectively). The value was almost on target when covariates were both generated from the Normal distribution but fell short when covariates were of mixed type (e.g., for a target median equal to 150 with $v = 2$, under random censoring, median survival times ranged from 131.0 to 156.5 and 93.0 to 114.0 when the two covariates were both normally distributed and mixed, respectively) (Figure 3, middle row).

Computational efficiency was affected by the choice of target median. The median run time increased as the target median increased (Figure 4). For random, administrative, and traditional censoring, respectively, run times ranged from 10.7 to 132.9 seconds, from 20.9 to 163.2 seconds, and from 49.7 to 213.2 seconds.

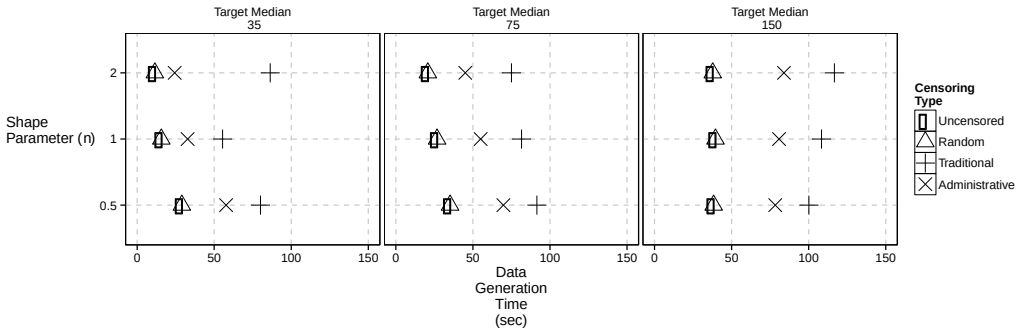


Figure 4. Median run times for censoring type when survival times are generated bounded between 20 and 300 and using mixed covariates

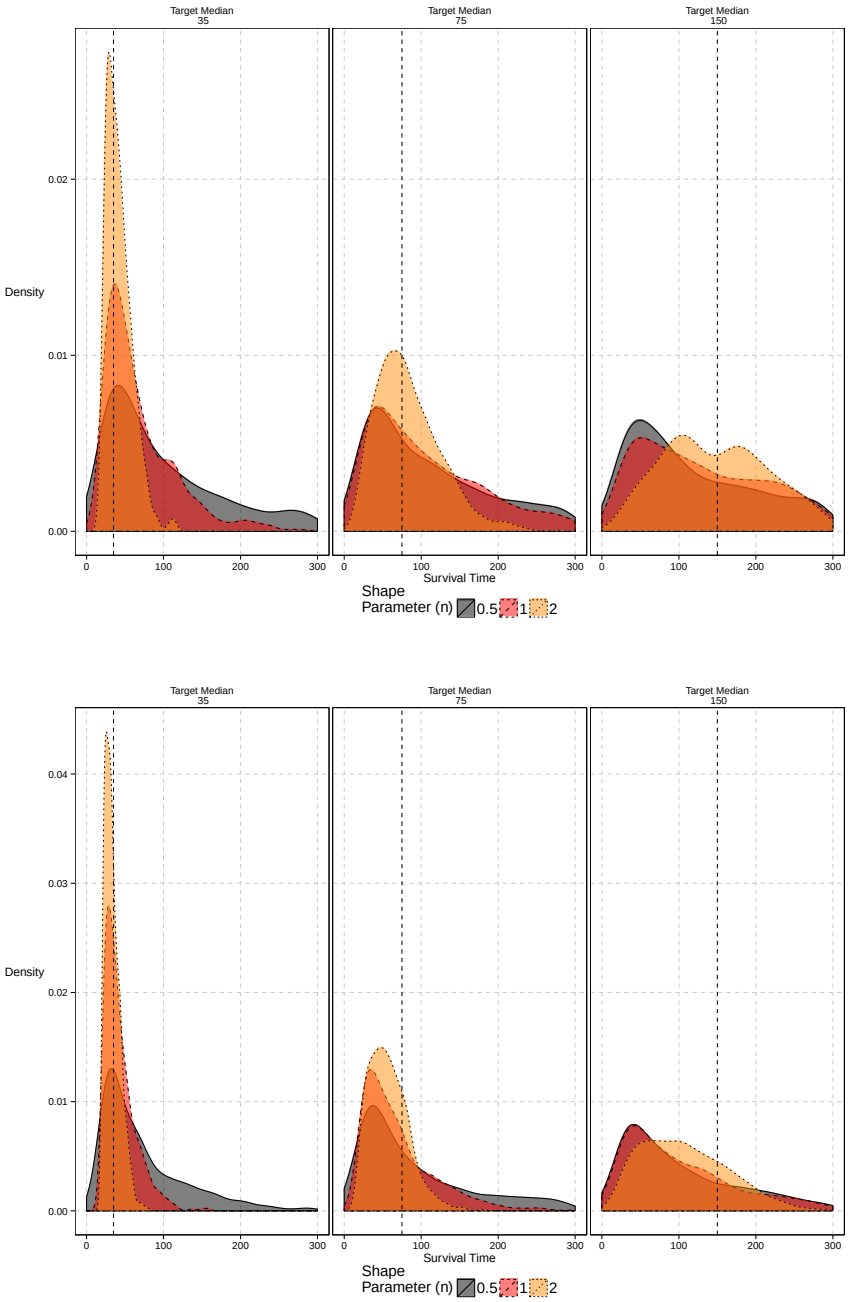


Figure 5. Example of density plots of generated survival times with random censoring for the various shape parameters when covariates are (a), above, two Normal random variables and (b), below, one Normal and one Bernoulli random variable

Overall, using a shape parameter equal to 2 provided distributions of survival times that have a median closer to the target value (Figure 5). We observed no differences in bias, MSE, or coverage probabilities between the choices of target median when the range of survival times is large (Figure S2).

Impact of Censoring Type

There was no difference in overall statistical performance of the algorithm by censoring type. No differences were found in standardized bias and coverage. For all censoring types, as the percent of censored observations increased, the MSE increased, ranging from 0.004 to 0.027 for the binary covariate and remaining low (on the order of 10-5) for Normal covariates (Figure S2).

However, it was found that computational run times were strongly affected by censoring type. Beyond the first step in the algorithm of generating uncensored survival times, random censoring took no additional time whereas traditional censoring more than doubled the run time (Figure 4). For example, for data generated with limits of 20 and 300, a target median of 75, and $v = 2$, median run times were 26.4, 26.6, and 89.2 seconds for uncensored, random and traditional censoring, respectively.

Impact of Type of Covariates and of Correlation

Negligible differences were found in performance by type of covariates or assumed correlation. In general, positive bias was found in the fitted coefficients corresponding to the binary covariate (e.g., for a target median equal to 35 and under random censoring, bias ranged from -4.0 to 9.7, -2.2 to 16.5, and 14.2 to 45.1 for bounds 20-300, 20-150, and 20-50, respectively) and negative bias for coefficients of the Normal covariates (e.g., for a target median equal to 35 and under random censoring, bias ranged from -12.3 to 2.1, -9.7 to -0.7, and -48.4 to -26.2 for bounds 20-300, 20-150, and 20-50, respectively). However, bias was negligible when the range of survival times generated (bounds) is large (Figure 1, left vs. right columns).

Median survival times generated were lower when using 1 Normal and 1 binary covariate compared to when both covariates were normally distributed (Figure 3). For example, for data generated between limits of 20 and 300 with a target median of 75 and $v = 2$, median survival times ranged from 36.0 to 60.0 and from 48.0 to 85.0 when the covariates were of mixed type and normally distributed, respectively (Figure 3, middle column).

In the subset of simulations performed to allow for varying correlation between the 2 covariates, we found a slight increase in the absolute MSE (e.g., from 2.0×10^{-5} to 4×10^{-5} for a Normal covariate with effect size equal to 0.02) as correlation increased. This is true for both combinations of the covariates (Figure 6). The MSE increased as the effect size increased, an effect that was more pronounced for the binary covariate (Figure 6b). For example, for data generated between 20 and 300, assuming a target median of 150, $\nu = 2$, and 50% observations randomly censored, when the two covariates were independently generated from a Normal distribution, the MSE for the covariate with effect size equal to 0.02 was 2.0×10^{-5}. When the effect size was instead 0.04, the MSE was 8.2×10^{-5}. The MSE for the coefficient of the uncorrelated binary covariate with an effect size of -0.5 equaled to 9.5×10^{-3}.

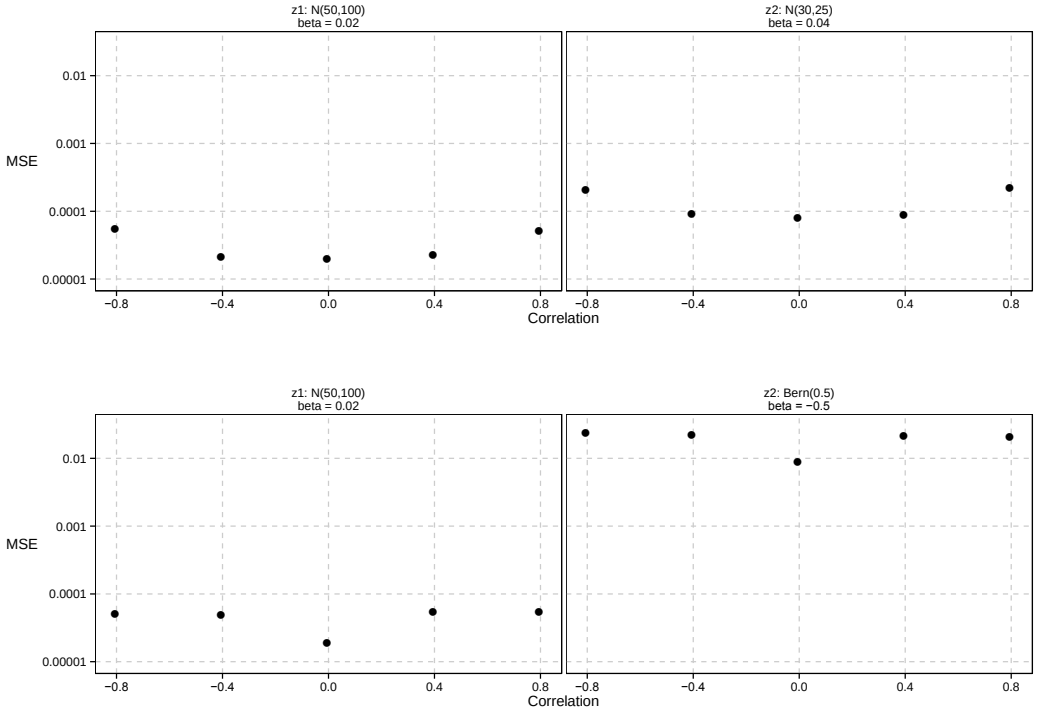


Figure 6. MSE of fitted coefficients by correlation amount between covariates when survival times are generated bounded between 20 and 300 with random censoring and 50% of observations censored when covariates are (a), above, two Normal random variables and (b), below, one Normal and one Bernoulli variable

Additional results in the Supplemental materials are shown with stratification on either censoring type or statistical performance metric to aid visual interpretation. Figures 1 and 2 are subsets of supplemental Figures S1 and S2. Tables with the information contained in Figures 1 to 4 are included in the supplemental material (Tables S1 to S4, respectively).

Discussion

Zhou's method (Zhou, 2001) of generating right-censored outcomes has been implemented by Hendry (2014) using the piecewise exponential framework and allows for an arbitrary number and functional form of the covariates. The main point of this study was to provide concrete recommendations for researchers interested in generating survival data with a specific structure in mind, as in mimicking a motivating data set from a real study. The algorithm proposed by Hendry offers flexibility, but the author did not provide guidance on how to choose parameters that will lead to data with desired features. In particular, one step of the algorithm requires the practitioner to choose an arbitrary monotone increasing function, g, such that $g(0) = 0$, and $g^{-1}(t)$ is differentiable. It was demonstrated that choosing a Weibull distribution for g(.) leads to a simple calculation that allows the practitioner to specify a target median survival time. This recommendation has important implications for practical use because it allows researchers to have much greater control over the generated data.

The simulation results show that, to minimize bias in fitted coefficients and achieve a realistic distribution of survival times, generating data with wider limits are better than keeping the range small even if the target median survival time is low. When generating data to achieve a target median survival time of 35, the standardized bias was high when survival times were generated between 20 and 50, but no meaningful bias was found if the range was expanded to 20-150 or 20-300. It was found, unexpectedly, that when using an overly-restrictive survival time interval with traditional censoring, bias was reduced as the amount of censoring increased (Figure 2). It is generally expected that higher percentages of censoring observations will either increase or have no effect on bias. Here, because the specified range of survival times was too restrictive, when applying traditional censoring we get an inverse relationship between the percentage of observations censored and bias.

This counterintuitive relationship is caused by the survival time generation algorithm's use of resampling to produce only survival times that fall within a specified interval. Consider the set of lower-risk individuals whose covariates

compel them to have an event later than the upper bound of the specified interval. The algorithm will use their data to repeatedly generate survival times until a time is produced that falls within the interval. From a modeling perspective, these lower-risk individuals are indistinguishable from the higher-risk individuals whose covariates compel them to have an event near the boundary of the specified time interval. The lower risk represented by these people's covariates is not reflected in their survival times, and the result is the bias we see with the 20-50 interval.

Traditional censoring affects the lower-risk subjects in a sample; subjects with survival times closer to the upper bound are more likely to be censored under traditional censoring than under random censoring. Subjects with an event near the upper bound of the pre-specified limits can be divided into two groups: 1) subjects with risk consistent with having an event near the boundary, or 2) subjects forced by the algorithm to have an event near the boundary despite their lower-risk covariates. Group 2's survival times are not indicative of the actual risk present in the covariates. Consequently, their inclusion in models estimating associations between covariates and risk results in bias. Traditional censoring removes members of group 2 at a proportionally higher rate than that of subjects whose survival times better reflect actual risk (group 1). The reduction in bias with increased percentages of subjects censored when using traditional censoring is caused by traditional censoring disproportionately removing the bias-causing portion of our sample. This bias in the observed sample produces the observed bias in the fitted coefficients when the range of possible survival times (bounds) is too restrictive. Random censoring targets all subjects equally, thus leaving the bias-producing component of our subjects proportionally intact, and so has a much less pronounced effect on bias.

Given these results, it is recommended that the range be wide enough to generate a distribution with the correct shape, but that it should not be too large to preserve reasonable computational efficiency. Because the algorithm generates survival times as a function of time-varying covariates that vary at integer-valued steps of the time scale, each subject will have as many records as the survival time generated. Thus, a large range means that some subjects will have many records. This is an important computational consideration, especially if generating data is just a first step in a much larger simulation.

There is a fine balance between the survival function, $\exp(\boldsymbol{\beta}\mathbf{Z}_j')$, and the baseline hazard, $h_0(t)$, that will influence the final survival times generated and the computational run times. Some of it can be controlled while defining the g function, which is a key component of the algorithm. It is suggested defining the g

function via a Weibull distribution with parameters informed by an empirical data set. One can look at the distribution of survival times to decide whether the g function should reflect an increasing, decreasing or constant baseline hazard, imposing a value for the shape parameter (v).

Also, the empirical median survival time can be used as the target median in the generated times and use this target median in a formula to compute the scale parameter (λ). It was found that the choice of g function worked well. Statistical performance did not vary by v or target median. However, we did find increased run times with increased target medians. Of note, this approach to generating survival times is still useful even if the observed survival data does not follow a Weibull distribution but the goal of the simulation study is to evaluate the performance of the Cox model. However, it might not be appropriate if, for example, the researcher's aim is to performance a power analysis.

The algorithm performed similarly for both combinations of covariates, but we found lower median survival times in the case of covariates of mixed type compared to the case of two Normal covariates. So, in order to achieve the target median when using covariates of mixed type, the values of the parameters needed to compute the scale parameter might need to be changed iteratively, mainly by inflating the target median, until the distribution of generated times adequately resembles the empirical target distribution.

There were no major issues when covariates were correlated. An increase was noted in the MSE, which was likely associated with an increase in the effect size and not necessarily with the type of covariates being used. Effect sizes play an important role as they have a direct impact on the distribution of the survival times. Further investigation is needed as well as exploring the performance of the algorithm in a scenario where correlations are observed within an individual.

Given the results shown by the simulation study, the following is suggested:

1. The use of the Weibull distribution to define the g function:
   $g = (\lambda^{-1}t)^{1/v}$ and $g^{-1} = \lambda t^{v}$
2. Parameters for the Weibull distribution can be informed from an empirical dataset:
   a. Use the distribution of survival times to decide if the g function should reflect an increasing, decreasing or constant baseline hazard to define the shape parameter (v);
   b. Use the observed median survival times to define a target median (M);

c. For a vector of effect estimates $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_p)$ and a vector of means of the covariates $\bar{\mathbf{Z}} = \left(\bar{Z}_1,\ldots,\bar{Z}_p\right)$, the scale parameter λ can be defined as follows:

$$\lambda = \frac{\log 2}{\boldsymbol{\beta}\bar{\mathbf{Z}}'} M^{-\nu}$$

3. Iterate until appropriate values can be found for the survival times. A wider range will yield a higher number of records per subject increasing the computational time. In contrast, a more limited range may introduce bias;
4. Utilize random censoring.

In conclusion, Hendry's algorithm for computing survival times that follow an extended Cox model with time-varying covariates were found to be a reasonable and practical solution when generating studies intended to closely resemble a motivating data set. Guidelines, substantiated by the simulation study, are provided to make this process easier.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Patient-Centered Outcomes Research Institute or the United States government.

Supplemental Material

Supplemental tables and figures are available in the Supplemental Material file, available at https://doi.org/10.22237/jmasm/1493597100.

References

Abrahamowicz, M., MacKenzie, T., & Esdaile, J. M. (1996). Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association, 91*(436), 1432-1439. doi: 10.2307/2291569

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine, 31*(29), 3946-3958. doi: 10.1002/sim.5452

Bakoyannis, G., & Touloumi, G. (2012). Practical methods for competing risks data: A review. *Statistical Methods in Medical Research, 21*(3), 257-272. doi: 10.1177/0962280210394479

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine, 24*(11), 1713-1723. doi: 10.1002/sim.2059

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*(24), 4279-4292. doi: 10.1002/sim.2673

Collett, D. (2003). *Modelling survival data in medical research* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. doi: 10.1037//1082-989x.6.4.330

Crowther, M. J., & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine, 32*(23), 4118-4134. doi: 10.1002/sim.5823

Demirtas, H., & Doganay, B. (2012). Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics, 22*(2), 223-236. doi: 10.1080/10543406.2010.521874

Desai, M., Bryson, S. W., & Robinson, T. (2013). On the use of robust estimators for standard errors in the presence of clustering when clustering membership is misspecified. *Contemporary Clinical Trials, 34*(2), 248-256. doi: 10.1016/j.cct.2012.11.006

Hendry, D. J. (2014). Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers. *Statistics in Medicine, 33*(3), 436-454. doi: 10.1002/sim.5945

Leemis, L. M., Shih, L.-H., & Reynertson, K. (1990). Variate generation for accelerated life and proportional hazards models with time dependent covariates. *Statistics & Probability Letters, 10*(4), 335-339. doi: 10.1016/0167-7152(90)90052-9

Sylvestre, M. P., & Abrahamowicz, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine, 27*(14), 2618-2634. doi: 10.1002/sim.3092

Zhou, M. (2001). Understanding the Cox regression models with time-change covariates. *The American Statistician, 55*(2), 153-155. doi: 10.1198/000313001750358491

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 107-126.
doi: 10.22237/jmasm/1493597160

ISSN 1538 − 9472

A Schmid-Leiman-Based Transformation Resulting in Perfect Inter-correlations of Three Types of Factor Score Predictors

André Beauducel
University of Bonn
Bonn, Germany

Factor score predictors are computed when individual factor scores are of interest. Conditions for a perfect inter-correlation of the best linear factor score predictor, the best linear conditionally unbiased predictor, and the determinant best linear correlation-preserving predictor are presented. A transformation resulting in perfect correlations of the three predictors is proposed.

Keywords: Factor analysis, factor score predictors, Schmid-Leiman transformation

Introduction

Because factor scores are not determinate (Guttman, 1955), they cannot be unambiguously computed. However, factor score predictors can be computed as linear combinations of the observed variables in order to represent the individual scores of a latent variable. This might be useful when decisions have to be justified on the individual score level. Several different factor score predictors have meanwhile been proposed (Mulaik, 2010). The properties of different factor score predictors have been investigated by means of simulation studies (Fava & Velicer, 1992) and by means of algebraic considerations (e.g. Beauducel & Hilger, 2015; Krijnen, 2006; Krijnen, Wansbeek & Ten Berge, 1996; McDonald & Burr, 1967; Schneeweiss & Mathes, 1995).

According to Grice (2001) and according to Krijnen et al. (1996) there are three main types of factor score predictors: The best linear predictor that is also known as Thurstone's (1935) regression predictor, the conditionally unbiased predictor (Krijnen et al., 1996; Bartlett, 1937), and the correlation-preserving predictor (McDonald, 1981; Ten Berge, Krijnen, Wansbeek, & Shapiro, 1999).

André Beauducel is Faculty in the Institute of Psychology. Email at beauducel@uni-bonn.de .

These three types of factor score predictors represent three desired properties: (a) The best linear predictor has a maximal correlation with the corresponding factor, (b) the conditionally unbiased predictor has zero correlations with non-corresponding factors, and (c) the correlation-preserving predictor has the advantage of preserving the correlations between the factors in the factor score predictor. The terms 'best linear predictor', 'conditionally unbiased predictor', and 'correlation-preserving predictor' are used as in Krijnen (2006).

McDonald and Burr (1967) explored the conditions for high correlations between factor score predictors for corresponding factors. They investigated the best linear predictor, a conditionally unbiased predictor, and a correlation preserving predictor. Since the determinant best linear correlation-preserving predictor (Ten Berge, Krijnen, Wansbeek, & Shapiro, 1999) was not available at that time, they explored the Anderson-Rubin's (1956) orthogonal (orthogonality preserving) factor score predictor. They found that the three factor score predictors are perfectly correlated for the one-factor model (the Spearman case). The investigated factor score predictors are perfectly correlated in the case of unrotated canonical factor analysis (Rao, 1955). McDonald and Burr (1967) acknowledged the preference to use rotated factor loadings, because they can often be interpreted more easily. However, for the rotated factors the correlations between the factor score predictors would generally not be perfect, leading to the problem of choosing the optimal factor score predictor.

There are at least three types of factor score predictors corresponding to three different desired properties (Grice, 2001). Moreover, there are conditions for which the correlations between the factor score predictors are one for corresponding factors, so that no choice has to be made (McDonald & Burr, 1967). It can be regarded as a substantial advantage of factor score predictors when they are simultaneously the best linear predictor, conditionally unbiased, as well as correlation preserving. Therefore, the aim of the present paper is (1) to explore further the conditions for perfect correlations between the factor score predictors of corresponding factors and (2) to propose a transformation method based on Schmid-Leiman (1957) that allows to find interpretable factors with perfect correlations between the three different types of factor score predictors.

Methodology

In order to present the equations defining the three factor score predictors, the definition of the population common factor model is given. The common factor

model assumes that **x**, the random vector of observations of order *p*, is generated by

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{e} \tag{1}$$

where **f** is the random vector of factor scores of order *q*, **e** the random error vector of order *p*, and **Λ** the factor pattern matrix of order *p* by *q*. The observations **x**, the factor scores **f**, and the error vectors **e** are assumed to have an expectation zero ($\varepsilon[\mathbf{x}] = 0$, $\varepsilon[\mathbf{f}] = 0$, $\varepsilon[\mathbf{e}] = 0$). The covariance between the factor scores and the error scores is assumed to be zero ($\mathrm{Cov}[\mathbf{f},\mathbf{e}] = 0$). The standard deviation of **f** is one, the covariance of the observed variables is $\mathbf{xx}' = \mathbf{\Sigma}$. The covariance matrix **Σ** can be decomposed by

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Psi}^2, \tag{2}$$

where **Φ** represents the *q* by *q* factor correlation matrix and $\mathbf{\Psi}^2$ the *p* by *p* covariance matrix of the error scores **e** ($\mathrm{Cov}[\mathbf{e},\mathbf{e}] = \mathbf{\Psi}^2$). $\mathbf{\Psi}^2$ is assumed to be a diagonal matrix and it will be assumed in this paper that the matrix is positive definite.

The regression predictor or best linear (BL) predictor is given by $\hat{\mathbf{f}}_{BL} = \mathbf{\Phi\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{x}$. The condition $\mathbf{B}'\mathbf{\Lambda} = \mathbf{I}$ holds for the class of conditionally unbiased predictors, where **B** are the weights for the factor score predictor (Bartlett, 1937). According to Krijnen et al. (1996), the best linear conditionally unbiased (BLCU) predictor is $\hat{\mathbf{f}}_{BCLU} = (\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{x}$. Ten Berge et al. (1999) defined a determinant best linear correlation-preserving (DBLCP) predictor, given by $\hat{\mathbf{f}}_{DBLCP} = \mathbf{\Phi}^{½}(\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda\Phi}^{½})^{-½}\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{x}$. For this predictor symmetric positive (semi) definite matrices are raised to a certain power (e.g. square-root) by raising its eigenvalues to that power. When the power of the eigenvalues is ½, this procedure is sometimes called the symmetric square-root (Harman, 1976).

Results

Conditions for a perfect correlation between $\hat{\mathbf{f}}_{BL}$, $\hat{\mathbf{f}}_{BLCU}$, and $\hat{\mathbf{f}}_{DBLCP}$

The following Theorem 1 to 3 describe the conditions for perfect correlations between the factor score predictors for corresponding orthogonal factors. As will

be shown in Theorem 4, a perfect correlation between the factor score predictors can only be found under unrealistic conditions when the factors are correlated. This is, of course, a limitation. However, the following Theorem 1 to 3 can nevertheless be applied to correlated factor solutions because correlated factor models can be transformed into corresponding orthogonal Schmid-Leiman (1957) models, as will be soon discussed.

Theorem 1 provides a condition for a perfect correlation between $\hat{\mathbf{f}}_{\text{BCLU}}$ and $\hat{\mathbf{f}}_{\text{BL}}$ for corresponding orthogonal factors.

Theorem 1. If $\boldsymbol{\Phi} = \mathbf{I}$ and $\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})$ then
$\varepsilon[\hat{\mathbf{f}}_{\text{BCLU}}\,\hat{\mathbf{f}}'_{\text{BL}}]\,\text{diag}(\varepsilon[\hat{\mathbf{f}}_{\text{BCLU}}\,\hat{\mathbf{f}}'_{\text{BLCU}}])^{-½}\,\text{diag}(\varepsilon[\hat{\mathbf{f}}_{\text{BL}}\,\hat{\mathbf{f}}'_{\text{BL}}])^{-½} = \mathbf{R}_{\text{BLCU,BL}} = \mathbf{I}.$

Proof. The covariance between $\hat{\mathbf{f}}_{\text{BCLU}}$ and $\hat{\mathbf{f}}_{\text{BL}}$ is

$$\mathbf{C}_{\text{BLCU,BL}} = \left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi} = \boldsymbol{\Phi}. \tag{3}$$

The correlation between $\hat{\mathbf{f}}_{\text{BCLU}}$ and $\hat{\mathbf{f}}_{\text{BL}}$ is therefore

$$\mathbf{R}_{\text{BLCU,BL}} = \boldsymbol{\Phi}\,\text{diag}\left(\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right)^{-½}\text{diag}\left(\boldsymbol{\Phi}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\right)^{-½}. \tag{4}$$

The element-wise square-root is calculated for the diagonal elements in Equation 4.

For $\boldsymbol{\Phi} = \mathbf{I}$ and $\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})$, Equation 4 can be transformed into

$$\mathbf{R}_{\text{BLCU,BL}} = \text{diag}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-½}\text{diag}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-½} = \mathbf{I}. \tag{5}$$

This completes the proof. □

The condition expressed in Theorem 1 is also a basis for a perfect correlation between $\hat{\mathbf{f}}_{\text{DBLCP}}$ and $\hat{\mathbf{f}}_{\text{BL}}$.

Theorem 2. If $\mathbf{\Phi} = \mathbf{I}$ and $\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda} = \text{diag}(\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda})$ then

$$\varepsilon[\hat{\mathbf{f}}_{\text{DBLCP}}\,\hat{\mathbf{f}}'_{\text{BL}}]\text{diag}(\varepsilon[\hat{\mathbf{f}}_{\text{DBLCP}}\,\hat{\mathbf{f}}'_{\text{DBLCP}}])^{-½}\,\text{diag}(\varepsilon[\hat{\mathbf{f}}_{\text{BL}}\,\hat{\mathbf{f}}'_{\text{BL}}])^{-½} = \mathbf{R}_{\text{DBLCP,BL}} = \mathbf{I}.$$

Proof. The covariance between $\hat{\mathbf{f}}_{\text{DBLCP}}$ and $\hat{\mathbf{f}}_{\text{BL}}$ is

$$\begin{aligned}\mathbf{C}_{\text{DBLCP,BL}} &= \mathbf{\Phi}^{½}\left(\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi}^{½}\right)^{-½}\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi} \\ &= \mathbf{\Phi}^{½}\left(\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi}^{½}\right)^{½}\mathbf{\Phi}^{½}.\end{aligned} \tag{6}$$

The corresponding correlation is

$$\mathbf{R}_{\text{DBLCP,BL}} = \mathbf{\Phi}^{½}\left(\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi}^{½}\right)^{½}\mathbf{\Phi}^{½}\text{diag}\left(\mathbf{\Phi}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi}\right)^{-½}. \tag{7}$$

For $\mathbf{\Phi} = \mathbf{I}$ and $\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda} = \text{diag}(\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda})$ Equation 7 can be transformed into

$$\mathbf{R}_{\text{DBLCP,BL}} = \left(\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\right)^{½}\left(\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\right)^{-½} = \mathbf{I}, \tag{8}$$

because the symmetric square-root and the conventional square-root are identical for diagonal matrices. This completes the proof. □

Finally, the condition presented in Theorem 1 and 2 is also the basis for a perfect correlation between $\hat{\mathbf{f}}_{\text{BCLU}}$ and $\hat{\mathbf{f}}_{\text{DBLCP}}$ for corresponding orthogonal factors.

Theorem 3. If $\mathbf{\Phi} = \mathbf{I}$ and $\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda} = \text{diag}(\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda})$ then

$$\varepsilon[\hat{\mathbf{f}}_{\text{BCLU}}\,\hat{\mathbf{f}}'_{\text{DBLCP}}]\text{diag}(\varepsilon[\hat{\mathbf{f}}_{\text{BCLU}}\,\hat{\mathbf{f}}'_{\text{BLCU}}])^{-½}\,\text{diag}(\varepsilon[\hat{\mathbf{f}}_{\text{DBLCP}}\,\hat{\mathbf{f}}'_{\text{DBLCP}}])^{-½} = \mathbf{R}_{\text{BLCU,DBLCP}} = \mathbf{I}.$$

Proof. The covariance between $\hat{\mathbf{f}}_{\text{BCLU}}$ and $\hat{\mathbf{f}}_{\text{DBLCP}}$ is

$$\begin{aligned}\mathbf{C}_{\text{BLCU,DBLCP}} &= \left(\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi}^{½}\left(\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi}^{½}\right)^{-½}\mathbf{\Phi}^{½} \\ &= \Phi^{½}\left(\mathbf{\Phi}^{½}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi}^{½}\right)^{-½}\mathbf{\Phi}^{½}.\end{aligned} \tag{9}$$

The corresponding correlation is

$$\mathbf{R}_{\text{BLCU,DBLCP}} = \boldsymbol{\Phi}^{½}\left(\boldsymbol{\Phi}^{½}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}^{½}\right)^{-½}\boldsymbol{\Phi}^{½}\text{diag}\left(\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}\right)^{-½}. \tag{10}$$

If $\boldsymbol{\Phi} = \mathbf{I}$ and $\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})$ Equation 10 can be transformed into

$$\mathbf{R}_{\text{BLCU,DBLCP}} = \left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-½}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{½} = \mathbf{I}. \tag{11}$$

This completes the proof. □

Thus, the correlations between $\hat{\mathbf{f}}_{\text{BL}}$, $\hat{\mathbf{f}}_{\text{BCLU}}$ and $\hat{\mathbf{f}}_{\text{DBLCP}}$, for corresponding orthogonal factors have been investigated for $\boldsymbol{\Phi} = \mathbf{I}$ and $\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})$. It turned out $\hat{\mathbf{f}}_{\text{BL}}$, $\hat{\mathbf{f}}_{\text{BCLU}}$ and $\hat{\mathbf{f}}_{\text{DBLCP}}$ are perfectly correlated for corresponding orthogonal factors with $\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})$. Therefore, the interesting properties of these three types of factor score predictors can be obtained by a single set of factor score predictors under the conditions expressed in Theorems 1, 2, and 3.

Theorem 4 shows that it is possible to get a perfect correlation $\hat{\mathbf{f}}_{\text{BL}}$ and $\hat{\mathbf{f}}_{\text{BCLU}}$, for the correlated factors model, if at least some observed variables are measured without error.

Theorem 4. If $\boldsymbol{\Phi} \neq \text{diag}(\boldsymbol{\Phi})$ then $\text{diag}(\mathbf{R}_{\text{BLCU,BL}}) = \mathbf{I}$ if $(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda})^{-1} = \mathbf{0}$.

Proof. From Jöreskog (1969; Equation 10) we get $\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda}(\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda})^{-1} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$. Entering $\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda}(\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda})^{-1}$ for $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$ into Equation 4 and some transformation yields

$$\text{diag}\left(\mathbf{R}_{\text{BLCU,BL}}\right) = \text{diag}\begin{pmatrix} \boldsymbol{\Phi}\ \text{diag}\left(\left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda}\right)^{-1} + \boldsymbol{\Phi}\right)^{-½} \\ \text{diag}\left(\boldsymbol{\Phi}\left(\left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda}\right)^{-1} + \boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}\right)^{-½} \end{pmatrix}. \tag{12}$$

For $(\mathbf{\Lambda'\Psi}^{-2}\mathbf{\Lambda})^{-1} = \mathbf{0}$, Equation 12 yields

$$\operatorname{diag}\left(\mathbf{R}_{\mathrm{BLCU,BL}}\right) = \operatorname{diag}\left(\mathbf{\Phi}\ \operatorname{diag}\left(\mathbf{\Phi}\right)^{-½}\ \operatorname{diag}\left(\mathbf{\Phi}\right)^{-½}\right) = \mathbf{I}. \tag{13}$$

This completes the proof. □

The condition $(\mathbf{\Lambda'\Psi}^{-2}\mathbf{\Lambda})^{-1} = \mathbf{0}$ can only be true if at least one observed variable of each factor is measured without error (Beauducel & Hilger, 2015). This is, however, not realistic and it was therefore excluded in the definition of the factor model that $\mathbf{\Psi}$ contains zero elements. Although it cannot be excluded that some transformation methods might be found that allow to find correlated factor models with perfect correlations between $\hat{\mathbf{f}}_{\mathrm{BL}}$ and $\hat{\mathbf{f}}_{\mathrm{BCLU}}$, Theorem 4 demonstrates that this is impossible with conventional properties of $(\mathbf{\Lambda'\Psi}^{-2}\mathbf{\Lambda})^{-1}$, which implies that the current approach is limited to orthogonal factor models. In order to overcome the limitation to orthogonal factor models Schmid-Leiman (1957) transformations of correlated factor models will be considered in the following.

Transformation resulting in perfect correlations between $\hat{\mathbf{f}}_{\mathrm{BL}}$, $\hat{\mathbf{f}}_{\mathrm{BLCU}}$, and $\hat{\mathbf{f}}_{\mathrm{DBLCP}}$

In the following, a transformation comprising four steps will be proposed that allows for orthogonal and correlated factors to be transformed into orthogonal (Schmid-Leiman) factors with perfect correlations between $\hat{\mathbf{f}}_{\mathrm{BL}}$, $\hat{\mathbf{f}}_{\mathrm{BCLU}}$ and $\hat{\mathbf{f}}_{\mathrm{DBLCP}}$. The transformation comprises four steps.

First, transform the factor loadings into

$$\mathbf{\Lambda}^{*} = \mathbf{\Lambda}\left(\mathbf{\Lambda'\Sigma}^{-1}\mathbf{\Lambda}\right)^{-½}\operatorname{diag}\left(\mathbf{\Lambda'\Sigma}^{-1}\mathbf{\Lambda}\right)^{½}. \tag{14}$$

It follows that

$$\begin{aligned}\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^{*} &= \operatorname{diag}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{½}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-½}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-½}\operatorname{diag}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{½} \\ &= \operatorname{diag}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right),\end{aligned} \tag{15}$$

which implies that $\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^{*} = \operatorname{diag}(\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^{*})$ holds for $\boldsymbol{\Lambda}^{*}$.

Second, calculate the factor inter-correlations $\boldsymbol{\Phi}^{*}$ for the corresponding loadings, because the transformation by means of Equation 14 modifies the factor inter-correlations as long as $\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \neq \mathbf{I}$, as follows from

$$\begin{aligned}\boldsymbol{\Phi} &\neq \boldsymbol{\Phi}^{*} \\ \left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\left(\boldsymbol{\Sigma}-\boldsymbol{\Psi}^{2}\right)\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1} &\neq \left(\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Lambda}^{*}\right)^{-1}\boldsymbol{\Lambda}^{*\prime}\left(\boldsymbol{\Sigma}-\boldsymbol{\Psi}^{2}\right)\boldsymbol{\Lambda}^{*}\left(\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Lambda}^{*}\right)^{-1} \\ \left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\left(\boldsymbol{\Sigma}-\boldsymbol{\Psi}^{2}\right)\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1} &\neq \left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-½}\operatorname{diag}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{½}\right)^{-1} \\ &\quad \boldsymbol{\Lambda}'\left(\boldsymbol{\Sigma}-\boldsymbol{\Psi}^{2}\right)\boldsymbol{\Lambda} \\ &\quad \left(\operatorname{diag}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{½}\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-½}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1}\end{aligned} \tag{16}$$

Thus, even when the initial factor model was orthogonal ($\boldsymbol{\Phi} = \mathbf{I}$), the transformed factor model will not necessarily be orthogonal ($\boldsymbol{\Phi}^{*} \neq \mathbf{I}$). As already noted, the transformation of the loadings according to Equation 14 can also be performed for correlated factors. It is, however, possible that $\operatorname{diag}(\boldsymbol{\Phi}^{*}) \neq \mathbf{I}$ as should be because $\boldsymbol{\Phi}^{*}$ is a correlation matrix (see definition of the factor model). In order to make sure that $\operatorname{diag}(\boldsymbol{\Phi}^{*}) = \mathbf{I}$ it is necessary to rescale $\boldsymbol{\Lambda}^{*}$ by means of $\boldsymbol{\Lambda}^{*}\operatorname{diag}(\boldsymbol{\Phi}^{*})^{-½}$ and to recalculate $\boldsymbol{\Phi}^{*}$ according to Equation 16. According to Theorems 1 to 4 it is, moreover, necessary to have orthogonal factors in order to get perfect correlations between $\hat{\mathbf{f}}_{\mathrm{BL}}$, $\hat{\mathbf{f}}_{\mathrm{BCLU}}$ and $\hat{\mathbf{f}}_{\mathrm{DBLCP}}$ for corresponding factors.
Third, perform a second order factor analysis so that

$$\boldsymbol{\Phi}^{*} = \boldsymbol{\Lambda}_{2}^{*}\boldsymbol{\Lambda}_{2}^{*\prime} + \boldsymbol{\Psi}_{2}^{*2}, \tag{17}$$

where the subscript denotes the parameters of the second order factor model.

Fourth, perform a Schmid-Leiman (1957) transformation in order to compute orthogonal primary factors. It is possible to perform a Schmid-Leiman transformation of more complex hierarchical models. However, in purpose of

brevity it is assumed here that $\mathbf{\Phi}^*$ can be decomposed into a single general (second order) factor and the corresponding uniqueness of the primary factors, that is

$$\mathbf{\Phi}^* = \mathbf{\Lambda}_2^* \mathbf{\Lambda}_2^{*\prime} + \mathbf{\Psi}_2^{*2} = \begin{bmatrix} \mathbf{\Lambda}_2^* & \vdots & \mathbf{\Psi}_2^{*2} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}_2^* \\ \mathbf{\Psi}_2^{*2} \end{bmatrix} = \mathbf{P}\mathbf{P}'. \tag{18}$$

The Schmid-Leiman transformation of the oblique first order factor model is

$$\mathbf{\Lambda}_{\mathrm{SL}}^* = \mathbf{\Lambda}^* \mathbf{P}. \tag{19}$$

It follows from Equations 2, 18, and 19 that

$$\mathbf{\Sigma} = \mathbf{\Lambda}^* \mathbf{\Phi}^* \mathbf{\Lambda}^{*\prime} + \mathbf{\Psi}^2 = \mathbf{\Lambda}_{\mathrm{SL}}^* \mathbf{\Lambda}_{\mathrm{SL}}^{*\prime} + \mathbf{\Psi}^2, \tag{20}$$

which implies that $\mathbf{\Lambda}_{\mathrm{SL}}$ represents the loadings of orthogonal factors. In the simplest Schmid-Leiman solution, the first column in $\mathbf{\Lambda}_{\mathrm{SL}}$ contains the loadings of the observed variables on a general (second order) factor that is orthogonal to the remaining orthogonalized primary factors.

However, the interest here is into the orthogonalized primary factors, which can be found in the columns 2 to *q*,

$$\mathbf{\Lambda}_{\mathbf{SLP}}^* = \begin{bmatrix} \lambda_{\mathbf{SL},1,2}^* & \cdots & \lambda_{\mathbf{SL},1,q}^* \\ \vdots & \ddots & \vdots \\ \lambda_{\mathbf{SL},p,2}^* & \cdots & \lambda_{\mathbf{SL},p,q}^* \end{bmatrix} \tag{21}$$

The subset of orthogonalized primary factors can also be calculated by means of

$$\mathbf{\Lambda}_{\mathbf{SLP}}^* = \mathbf{\Psi}_2^2 \mathbf{\Lambda}_2^*. \tag{22}$$

According to Equation 14 this implies

$$\begin{aligned} \mathbf{\Lambda}_{\mathbf{SLP}}^{*\prime} \mathbf{\Sigma}^{-1} \mathbf{\Lambda}_{\mathbf{SLP}}^* &= \mathbf{\Psi}_2^{*2} \mathbf{\Lambda}^{*\prime} \mathbf{\Sigma}^{-1} \mathbf{\Lambda}^* \mathbf{\Psi}_2^{*2} \\ &= \mathbf{\Psi}_2^{*2} \mathrm{diag}\left(\mathbf{\Lambda}' \mathbf{\Sigma}^{-1} \mathbf{\Lambda} \right) \mathbf{\Psi}_2^{*2} = \mathrm{diag}\left(\mathbf{\Lambda}_{\mathbf{SLP}}^{*\prime} \mathbf{\Sigma}^{-1} \mathbf{\Lambda}_{\mathbf{SLP}}^* \right), \end{aligned} \tag{23}$$

so that the conditions for perfect correlations of $\hat{\mathbf{f}}_{BL}$, $\hat{\mathbf{f}}_{BCLU}$ and $\hat{\mathbf{f}}_{DBLCP}$ are met for the corresponding orthogonalized primary factors.

Example

A correlation matrix presented by Rimoldi (1948) based on 19 ability tests assessed in 138 participants was used in order to illustrate the transformation described above. As an initial factor model, principal axis factoring of the correlation matrix with subsequent oblique rotation (Promax, kappa = 4) was performed with IBM SPSS Version 22 (see Table 1). The factor loading pattern and the factor inter-correlations were entered into the SPSS syntax presented in Appendix A in order to calculate the correlations between $\hat{\mathbf{f}}_{BL}$, $\hat{\mathbf{f}}_{BCLU}$ and $\hat{\mathbf{f}}_{DBLCP}$ for the corresponding factors of the initial factor model. Appendix A also contains the four steps of the procedure described before and can be adapted for other data sets when the corresponding loading pattern and factor inter-correlations as well as the number of second order factors for the Schmid-Leiman solution is entered.

As can be seen from Table 2 the correlations between $\hat{\mathbf{f}}_{BL}$, $\hat{\mathbf{f}}_{BCLU}$ and $\hat{\mathbf{f}}_{DBLCP}$ were already very high for the corresponding factors of the initial factor model. It should, however, be noted that the factor score predictors were based on exactly the same sample, the same observed variables and are thought to represent exactly the same factors. From this perspective especially some of the correlations between $\hat{\mathbf{f}}_{BL}$ and $\hat{\mathbf{f}}_{BCLU}$ indicate that the factor score predictors introduce a notable difference in the measurement of the same factors with the same participants. Therefore, a transformation of these factors according to the procedure described before was performed.

Table 1. Promax-rotated loading pattern and factor inter-correlations for 19 ability variables from Rimoldi (1948)

Variable	F1	F2	F3	F4	F5	F6	F7
1	-.02	**.56**	-.08	.01	.00	.12	.12
2	-.11	**.38**	**.33**	-.03	.06	.06	-.04
3	.03	.03	-.03	**.67**	.00	-.06	-.22
4	-.04	.18	**.54**	-.08	.11	-.25	.24
5	-.02	**.35**	.01	.20	.13	-.13	.06
6	-.01	.06	.16	.14	**.74**	-.07	**-.34**
7	.04	.02	.02	-.24	**.37**	.25	.29
8	.10	.15	.25	.25	-.05	-.11	.24
9	.06	.04	.00	-.09	-.19	-.06	**.47**
10	-.01	.18	-.08	.03	-.02	**.67**	-.09
11	-.03	-.13	-.02	**.59**	.16	.25	.15
12	.06	**.45**	.07	-.21	.17	.09	-.19
13	-.01	**.37**	.29	.03	-.07	.26	-.10
14	**.31**	-.10	**.65**	.01	.14	.02	-.17
15	**.65**	-.14	.28	.05	-.09	.10	.00
16	**.72**	-.12	.15	-.10	.08	-.07	.15
17	**.88**	.14	.00	.04	-.12	.02	-.01
18	**.59**	.17	-.26	.05	.19	-.05	.01
19	.08	**.43**	.13	.07	-.18	.05	.16
			factor inter-correlations				
F2	.37						
F3	.52	.22					
F4	.37	.18	.39				
F5	.28	.17	.25	.19			
F6	.02	-.12	.17	.02	.19		
F7	.29	.04	.35	.35	.52	.45	

Note. Loadings with an absolute size ≥ .30 are given in bold face.

Table 2. Correlations between $\hat{\mathbf{f}}_{BL}$, $\hat{\mathbf{f}}_{BLCU}$, and $\hat{\mathbf{f}}_{DBLCP}$ for the corresponding factors of the initial factor model

	F1	F2	F3	F4	F5	F6	F7
$\hat{\mathbf{f}}_{BL}$ with $\hat{\mathbf{f}}_{BLCU}$	.993	.986	.968	.983	.979	.984	.947
$\hat{\mathbf{f}}_{BL}$ with $\hat{\mathbf{f}}_{DBLCP}$	.999	.996	.992	.995	.995	.996	.987
$\hat{\mathbf{f}}_{BLCU}$ with $\hat{\mathbf{f}}_{DBLCP}$	.998	.997	.992	.996	.995	.996	.986

In the first step of the transformation described above, the factor loading pattern was transformed according to Equation 14 (see Appendix A). In the second step, the factor inter-correlations were calculated for the transformed loading pattern

(Equation 16). The loading pattern and the factor inter-correlations were rescaled. Third, an unrotated second order principal axis factoring of the inter-correlations of the factors was performed. A single second order factor was extracted. Fourth, a Schmid-Leiman solution was computed from the second order factor and the transformed primary factors (Equation 19; see Table 3). It turned out that the loading pattern of the initial primary factors and the loading pattern of the transformed Schmid-Leiman primaries were similar, which implies that the interpretation of the factors was not substantially altered by the transformations. The correlations between $\hat{\mathbf{f}}_{BL}$, $\hat{\mathbf{f}}_{BCLU}$ and $\hat{\mathbf{f}}_{DBLCP}$ for the corresponding primary factors presented in Table 3 were all perfect (= 1.000) so that an additional table was not necessary.

Table 3. Schmid-Leiman model of the primary factors transformed according to (14)

	2nd order factor	***Primary Factors***						
Variable	**F1**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**
1	.15	.05	**.52**	-.03	.04	.06	.09	.10
2	.16	.01	**.37**	.29	.02	.06	.05	.00
3	.06	.09	.07	.03	**.60**	-.02	-.10	-.12
4	**.30**	.11	.22	**.48**	.03	.15	-.17	.21
5	.14	.07	**.36**	.04	.21	.14	-.14	.07
6	.18	.10	.12	.14	.13	**.61**	-.08	-.13
7	**.33**	.08	.01	.05	-.17	.40	.29	**.30**
8	**.30**	.20	.19	.28	**.31**	.03	-.08	.21
9	.16	.07	.03	.03	-.02	-.08	-.01	**.33**
10	.15	-.01	.12	-.02	.00	.00	**.59**	-.01
11	**.36**	.05	-.11	.07	**.55**	.19	.25	.22
12	.03	.10	**.43**	.06	-.18	.14	.04	-.13
13	.17	.07	**.35**	.28	.06	-.05	.22	-.04
14	**.31**	**.39**	-.01	**.60**	.09	.14	.05	-.04
15	**.30**	**.61**	-.06	**.34**	.13	-.02	.11	.06
16	**.33**	**.67**	-.02	.23	.02	.16	-.03	.17
17	.26	**.80**	.23	.14	.13	-.02	.00	.02
18	.18	**.53**	.23	-.13	.09	.23	-.07	.05
19	.19	.15	**.42**	.17	.12	-.10	.04	.12

Note. Loadings with an absolute size ≥ .30 are given in bold face.

Conclusion

Conditions were explored for a perfect correlation between three types of factor score predictors: The regression predictor or best linear predictor, the conditionally unbiased best linear predictor, and the determinant best linear correlation-preserving predictor. A perfect correlation between these factor score predictors for corresponding factors implies that the choice between these factor score predictors does not matter and that each type of factor score predictor will have the virtues of the other. That is, the conditionally unbiased best linear predictor will also be the best linear predictor, the determinant best linear correlation-preserving predictor, will have the virtue to be conditionally unbiased predictor, etc. Thus, the conditions of a perfect correlation between the three types of factor score predictors for corresponding factors might be of interest for applied researchers, who want to calculate score predictors combining the different advantages.

McDonald and Burr (1967) found three types of factor score predictors similar to the predictors investigated here are perfectly correlated for one-factor models and for the unrotated canonical factor model. In addition to these conditions, it was shown here that for orthogonal factors with $\mathbf{\Lambda'\Sigma^{-1}\Lambda} = \text{diag}(\mathbf{\Lambda'\Sigma^{-1}\Lambda})$ the three factor score predictors are perfectly correlated. A method for transforming a loading matrix according to this condition was proposed. The transformation can also be applied to models with correlated factors. Moreover, the factors resulting from this transformation are not necessarily orthogonal. Since it has been shown that the factors corresponding to $\mathbf{\Lambda'\Sigma^{-1}\Lambda} = \text{diag}(\mathbf{\Lambda'\Sigma^{-1}\Lambda})$ should be orthogonal in order to provide perfect correlations between the three types of factor score predictors for corresponding factors a hierarchical Schmid-Leiman solution was computed. Thereby the correlated factor models are transformed into a combined solution of orthogonal second order factors and orthogonal primary factors. Since the Schmid-Leiman transformation can be applied to any hierarchical pattern of loading matrices, the transformation method proposed here can also be applied to confirmatory factor models.

The results of the current study show that it is possible to obtain a single set of factor score predictors that combine the virtues of the best linear predictor, of the conditionally unbiased predictor, and of the correlation-preserving predictor. This may be of interest for research and applications, where a high quality of the factors score predictors is of special importance.

As an example, the transformation was applied to the data set of Rimoldi (1948), who published the correlation matrix of 19 ability measures. The corresponding SPSS syntax (Appendix A) can be adapted in order to be used for other data sets.

References

Anderson, T. W. & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability, 5*, 111–150.

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology, 28*(1), 97–104. doi: 10.1111/j.2044-8295.1937.tb00863.x

Beauducel, A. & Hilger, N. (2015). Extending the debate between Spearman and Wilson 1929: When do single variables optimally reproduce the common part of the observed covariances? *Multivariate Behavioral Research, 50*(5), 555-567. doi: 10.1080/00273171.2015.1059311

Fava, J. L. & Velicer, W. F. (1992). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research, 27*(3), 301-322. doi: 10.1207/s15327906mbr2703_1

Guttman, L. (1955). The determinacy of factor score matrices with applications for five other problems of common factor theory. *British Journal of Statistical Psychology, 8*(2), 65-82. doi: 10.1111/j.2044-8317.1955.tb00321.x

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450. doi: 10.1037/1082-989X.6.4.430

Harman, H. H. (1976). *Modern Factor Analysis* (3rd ed.). Chicago: The University of Chicago Press.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202. doi: 10.1007/bf02289343

Krijnen, W. P. (2006). Some results on mean square error for factor score prediction. *Psychometrika, 71*(2), 395-409. doi: 10.1007/S 11336-004-1220-7

Krijnen, W. P., Wansbeek, T.J., & Ten Berge, J.M.F. (1996). Best linear predictors for factor scores. *Communications in Statistics: Theory and Methods, 25*(12), 3013–3025. doi: 10.1080/03610929608831883

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika, 46*(3), 337-341. doi: 10.1007/BF02293740

McDonald, R. P. & Burr, E. J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika, 32*(4), 381-401. doi: 10.1007/BF02289653

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd Ed.). New York: CRC Press.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika, 20*(2), 93-112. doi: 10.1007/BF02288983

Rimoldi, H. J. (1948). Study of some factors related to intelligence. *Psychometrika, 13*(1), 27-46. doi: 10.1007/BF02288945

Schmid, J. & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*(1), 53-61. doi: 10.1007/BF02289209

Schneeweiss, H. & Mathes, H. (1995). Factor Analysis and Principal Components. *Journal of Multivariate Analysis, 55*(1), 105-124. doi: 10.1006/jmva.1995.1069

Ten Berge, J. M. F., Krijnen, W. P., Wansbeek, T., Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and its Applications, 289*(1-3), 311-318. doi: 10.1016/S0024-3795(97)10007-6

Thurstone, L.L. (1935). *The Vectors of Mind*. Chicago: University of Chicago Press.

Appendix A

```
set MXLOOPS=1000 workspace=400000.


MATRIX.

* ENTER INITIAL LOADING PATTERN INTO L:.
compute L={
-0.019, 0.555,-0.083, 0.012, 0.005, 0.122, 0.123;
-0.108, 0.384, 0.334,-0.028, 0.064, 0.060,-0.042;
 0.033, 0.027,-0.033, 0.671, 0.002,-0.059,-0.220;
-0.035, 0.183, 0.541,-0.079, 0.105,-0.246, 0.244;
-0.024, 0.353, 0.007, 0.196, 0.129,-0.140, 0.063;
-0.014, 0.059, 0.164, 0.141, 0.744,-0.074,-0.343;
 0.044, 0.020, 0.021,-0.236, 0.371, 0.251, 0.291;
 0.098, 0.154, 0.250, 0.254,-0.046,-0.112, 0.236;
 0.064, 0.039,-0.005,-0.086,-0.192,-0.065, 0.470;
-0.008, 0.176,-0.079, 0.030,-0.022, 0.667,-0.089;
-0.030,-0.128,-0.021, 0.593, 0.163, 0.252, 0.146;
 0.058, 0.447, 0.067,-0.214, 0.170, 0.088,-0.190;
-0.012, 0.366, 0.290, 0.027,-0.073, 0.264,-0.100;
 0.315,-0.095, 0.647, 0.007, 0.146, 0.024,-0.170;
 0.648,-0.146, 0.280, 0.049,-0.087, 0.101, 0.001;
 0.719,-0.121, 0.152,-0.095, 0.075,-0.074, 0.153;
 0.879, 0.141, 0.002, 0.039,-0.117, 0.016,-0.012;
 0.586, 0.170,-0.265, 0.050, 0.186,-0.053, 0.011;
 0.076, 0.434, 0.128, 0.069,-0.176, 0.050, 0.164}.

* ENTER INITIAL FACTOR INTER-CORRELATIONS INTO PHI:.
compute Phi={
1.000, 0.366, 0.518, 0.372, 0.279, 0.017, 0.285;
0.366, 1.000, 0.219, 0.180, 0.170,-0.118, 0.035;
0.518, 0.219, 1.000, 0.385, 0.246, 0.173, 0.348;
0.372, 0.180, 0.385, 1.000, 0.192, 0.019, 0.353;
0.279, 0.170, 0.246, 0.192, 1.000, 0.186, 0.521;
0.017,-0.118, 0.173, 0.019, 0.186, 1.000, 0.449;
0.285, 0.035, 0.348, 0.353, 0.521, 0.449, 1.000}.
```

```
* ENTER NUMBER OF SECOND ORDER FACTORS.
compute nF_2nd=1.

compute Psi2=Mdiag(diag( ident(nrow(L),nrow(L)) - L*Phi*T(L)  )).
compute Sig=L*Phi*T(L) + Psi2.


Print /Title "Initial factor loading pattern:".
print {L}/format=F5.2.
Print /Title "Initial factor inter-correlations:".
print {Phi}/format=F5.2.
Print /Title "Number of factors for second order factor analysis:".
print nf_2nd/format=F2.0.


Print /Title "Initial correlation between BLCU and BL factor score predictor
(Equation 4):".
compute EQ4=Phi*INV(Mdiag(diag(INV(T(L)*INV(Sig)*L))))&**(0.5) *
                INV(Mdiag(diag(Phi*T(L)*INV(Sig)*L*Phi)))&**(0.5).
print EQ4/format=F6.3.

Print /Title "Initial correlation between DBLCP and BL factor score predictor
(Equation 7):".
CALL SVD(Phi, q, eig, qq).
compute Phi12=q*(eig&**0.5)*T(q).
compute H=Phi12*T(L)*INV(Sig)*L*Phi12.
CALL SVD(H, q, eig, qq).
compute H12=q*(eig&**0.5)*T(q).
compute EQ7=Phi12*H12*Phi12*INV(Mdiag(diag(Phi*T(L)*INV(Sig)*L*Phi)))&**(0.5).
print EQ7/format=F6.3.

Print /Title "Initial correlation between BLCU and DBLCP factor score predictor
(Equation 10):".
compute EQ10=Phi12*INV(H12)*Phi12*INV(Mdiag(diag(INV(T(L)*INV(Sig)*L))))&**(0.5).
print EQ10/format=F6.3.
```

```
* TRANSFORMATION OF PRIMARY FACTORS:.

* STEP 1 - Compute transformed loadings according to Equation 14.
compute help=  T(L)*INV(Sig)*L  .
CALL SVD(help, V, Eig, TV).
compute help12=V*(Eig&**0.5)*T(V).
compute L14= L  * INV(help12)*Mdiag(diag( T(L)*INV(Sig)*L ))&**0.5.


* STEP 2 - Compute factor intercorrelations and rescale transformed loadings.
compute Phi14=INV(T(L14)*L14)*T(L14)*L*Phi*T(L) *L14*INV(T(L14)*L14).
compute L14=L14*(Mdiag(diag(Phi14)))&**0.5.
Print /Title "STEP 1 + 2 - Loading pattern of rescaled transformed primary
factors:".
Print L14/format=F5.2.
compute Phi14=INV(T(L14)*L14)*T(L14)*L*Phi*T(L) *L14*INV(T(L14)*L14).
Print /Title "STEP 1 + 2 - Inter-correlations of transformed primary factors:".
Print Phi14/format=F5.2.


* STEP 3 - Principal Axis Factoring of the intercorrelations of the transformed
primary factors
  (second order factor analysis).
compute R=Phi14.
* Initial PCA.
CALL EIGEN(R, PC, PC_eig).
compute PC_eig=Mdiag(PC_eig).
compute PC=PC*(PC_eig&**0.5).
compute A=PC(:,1).
LOOP i=2 to nF_2nd.
compute A={A,PC(:,i)}.
END LOOP.
* EFA.
compute F=A.
LOOP ii=1 to 50.
compute Rrep=R - ident(nrow(A),nrow(A)) + Mdiag(diag(F*T(F)))  .
CALL EIGEN(Rrep, FF, F_eig).
compute F_eig=ABS(Mdiag(F_eig)).
```

```
compute FF=FF*(F_eig&**0.5).
compute F=FF(:,1).
LOOP i=2 to nF_2nd.
compute F={F,FF(:,i)}.
END LOOP.
END LOOP.
compute F=-1*F.

compute Psi=Mdiag(diag(Phi14-F*T(F)))&**0.5.
compute P={F,Psi}.
Print /Title "STEP 3 - 2nd order factor loadings with (diagonal) error factor
loadings:".
print {P} /format=F5.2.

* STEP 4 - Compute Schmid-Leiman solution.
compute SL14=L14*P.
Print /Title "STEP 4 - Schmid-Leiman Solution:".
print {SL14} /format=F5.2.



* CHECK: Compute the inter-correlations between factor score predictors for the
primaries.
* SELECT PRIMARIES OF SCHMID-LEIMAN SOLUTION:.
compute SL_p=SL14(:,2).
LOOP i=nF_2nd+2 to nF_2nd+ncol(L).
compute SL_p={SL_p,SL14(:,i)}.
END LOOP.

Print /Title "Correlation between BLCU and BL factor score predictor (Equation
4) for transformed primaries:".
compute EQ4_14=   INV(Mdiag(diag(GINV(T(SL_p)*INV(Sig)*SL_p))))&**(0.5) *
                 INV(Mdiag(diag(T(SL_p)*INV(Sig)*SL_p)))&**(0.5).
print EQ4_14/format=F6.3.

Print /Title "Correlation between DBLCP and BL factor score predictor (Equation
7) for transformed primaries:".
compute H=T(SL_p)*INV(Sig)*SL_p.
```

```
CALL SVD(H, q, eig, qq).
compute H12=q*(eig&**0.5)*T(q).
compute EQ7_14=H12*INV(Mdiag(diag(T(SL_p)*INV(Sig)*SL_p)))&**(0.5).
print EQ7_14/format=F6.3.

Print /Title "Correlation between BLCU and DBLCP factor score predictor
(Equation 10) for transformed primaries:".
compute EQ10_14=INV(H12)*INV(Mdiag(diag(INV(T(SL_p)*INV(Sig)*SL_p))))&**(0.5).
print EQ10_14/format=F6.3.

Print /Title
"Weights (B) for computation of factor scores as fscore=T(B)*Z, with Z "
+ "containing z-standardized variables with rows=variables, columns=cases".
compute B=INV(Sig)*SL_p.
print B/format=F6.3.

* For calculating the factor scores delete the first "*" in the three lines
starting with “get”.
* Enter the number of z-standardized variables in "##" and the file handle in
"...".
*get Z / variables= Z1 to Z## /file="C:\...\zscores.sav".
*compute Fscores=Z*B.
*save { Fscores } /outfile="C:\...\Fscores.sav".


END MATRIX.
```

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 127-136.
doi: 10.22237/jmasm/1493597220

ISSN 1538 − 9472

A Review of the Multiple-Sample Tests for the Continuous-Data Type

Dewi Rahardja
United States Department of Defense
Fort Meade, MD

For continuous data, various statistical hypotheses testing methods have been extensively discussed in the literature. In this article a review is provided of the multiple-sample continuous-data testing methods. It includes traditional methods, such as the two-sample *t*-test, Welch ANOVA test, etc., as well as newly-developed ones, such as the various Multiple Comparison Procedure (MCP). A roadmap is provided in a figure or diagram format as to which methods are available in the literature. Additionally, the implementation of these methods in popular statistical software packages such as SAS is also presented. This review will be helpful to determine which continuous-data testing method (along with the corresponding SAS code) are available to use in various fields of study, both for the design phase of a study in prospective study, cross-sectional, or retrospective study analysis and the analysis phase.

Keywords: Two-sample *t*-test, one-way ANOVA, Satterthwaite, degrees of freedom, Welch ANOVA, Wilcoxon rank-sum test, Kruskal-Wallis test, paired *t*-test, multiple comparison procedure (MCP)

Introduction

In many real-world applications, such as data in clinical trials, financial data, epidemiology, sociology, etc., we often encounter data with outcome (or response) variables that are continuous in nature. If a random variable can take any value within an interval or continuum, it is called a continuous random variable. For example, diastolic blood pressure, amount of dollar expenses, height, weight, cholesterol level, air pollutant level, etc. are usually considered continuous random variables because they can take any value within certain intervals, even though the observed measurement is limited by the accuracy of the measuring device. Due to the nice asymptotic math/stat properties, the Normal distribution is the most commonly-used continuous distribution in the fields of clinical research, finance, epidemiology, sociology, along with many others.

Dewi Rahardja is a Statistician. Email her at: rahardja@gmail.com.

Without loss of generality (WLOG), the standard (classical or frequentist) large sample (asymptotic) theory is derived using the underlying assumptions of independent, normally-distributed random variables with homogeneous (i.e., equal) variance.

A frequent task in data analysis is to check these three assumptions (in the order of: independent, normal, equal variance) for the outcome measure or response variable, and then to determine what test is suitable/appropriate for a dataset.

Such continuous-data outcome measure or response variables (or dependent variables) can occur both in randomized controlled trials and in observational studies. The predictor or covariate (or independent variable) is the terminology used for both continuous and categorical variable. However specifically, the predictor is called a grouping variable (or factor) for a discrete/categorical predictor. Typically, this grouping variable can have one, two, or multiple levels. The common (or generic) statistical terms used are one-, two-, and multiple-sample testing methods for one, two, and multiple levels of this one factor (or grouping variable).

To date, there is no literature that comprehensively presents and summarizes the review of the various one-sample, two-sample, and multiple-sample tests for the continuous-data type of response variable (or outcome measure) with one grouping variable (factor) of multiple levels. Hence in our line of (statistical) practice, we often find both statistician and non-statistician practitioners, investigators, and researchers get confused/mixed-up about the method, model, and hypothesis to use. To close this confusion gap, this article will be a very useful basic guidance/roadmap to both statisticians and non-statisticians in various fields of study.

For the categorical-data type (of outcome measure or response variable), Rahardja, Yang, and Zhang (2016) have provided a comprehensive review, also in a roadmap format, along with the corresponding translation/implementation of those methods in popular and professional statistical software packages, such as SAS and/or R.

Hypothesis Testing

First, we begin with the popular one-sample mean test (for a normal population): the one-sample *z*-test and the one-sample *t*-test (not listed on Table 1 nor Figure 1). WLOG, consider the simplest case: a continuous response variable (or outcome measure), Y, with one grouping variable (or factor), X, as the discrete

covariate or factor (or predictor). This single factor has only one level (i.e., $X = 1$). For this very basic/simplest model, the objective is to model the expected value of a continuous random variable, Y, as a linear function of the discrete predictor or factor, X, and hence $E(Y_i) = \mu_X$. This basic/simplest model has only one factor with one level (i.e., $X = 1$); therefore $E(Y_i) = \mu$. Hence, this (generic) model structure can be written as $Y_i = \mu_X + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, for $i = 1, 2,\ldots, n$ observations (which is a statistical linear model which is linear in the parameter, μ). Essentially, this model structure can be simplified as the mean model (for one factor), $Y_i = \mu + \epsilon_i$ where $i = 1, 2,\ldots, n$ observations. For this (generic) basic model the assumptions are that Y is normally distributed, errors are normally distributed and independent with constant/homogeneous variance σ^2, i.e. $\epsilon_i \sim N(0, \sigma^2)$, while X is fixed (i.e., $X = 1$); see Casella (2008).

Theoretically, with a known standard deviation (σ), the standard one-sample z-test can be used to test the null hypothesis, H_0: $\mu = 0$, versus the alternative hypothesis, H_1: $\mu \neq 0$. However, practically, the standard deviation (σ) is unknown, and hence the one-sample t-test can be used to test the same aforementioned hypothesis.

Second, consider the two-sample (and subsequently, multiple-sample) mean test (see Figure 1), depending on the assumptions of the response variable (or outcome measure). Consider the case: a continuous response variable (or outcome measure), Y, with one grouping variable (or factor), X, as the discrete covariate or factor (or predictor). This single factor has two (or more) levels (e.g., $X = 0$ for the placebo group, or for $X = 1$ the drug A group, or $X = 2$ for the drug B group, etc.), and can be written as an indicator function/variable. This model structure can be written as the so-called cell means model (for one factor), $Y_{ij} = \mu_i + \epsilon_{ij}$, where $i = 1, 2,\ldots, k$ groups (i.e., the i^{th} level of that one factor), and $j = 1, 2,\ldots, n$ observations; see Casella (2008). The model assumptions are that Yij is normally distributed, errors are normally distributed and independent with constant/homogeneous variance σ^2, i.e. $\epsilon_{ij} \sim N(0, \sigma^2)$; X is a fixed indicator function/variable (i.e., $X = 0, 1,$ etc.); and μ_i is the unknown theoretical/population mean for all of the observations at level i.

The generic hypothesis testing for two means can be written as H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, and for multiple means it can be generalized as H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ versus H_1: at least one mean is different than the rest.

Next, consider how to implement these methods (in Figure 1) in popular statistical software packages, such as SAS (see Table 1). The SAS PROC TTEST, or the TTEST procedure, performs t-tests for one-sample, two-sample, and paired observations (see Table 1 and Figure 1). The one-sample t-test compares the mean

of the sample to a given number (which you supply, and typically is zero). The dependent-sample or paired *t*-test compares the difference in the means from the two variables to a given number (usually 0) while taking into account the fact that the scores are not independent (i.e., paired scores or data); see David and Gunnink (1997). The independent samples *t*-test (or two-sample *t*-test) compares the difference in the means from the two groups to a given value (usually 0). In other words, it tests whether the difference in the means is 0.

When there are multiple levels within that one factor (or one way) model (of the cell means model), alternatively the model can be written as the effect model to test the effect of the multiple levels (i.e., multiple-sample test); similarly for the two levels (i.e., two-sample test). The effect model is used to separate the baseline mean effect from the groups' or levels' effect: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where $i = 1, 2,\ldots, k$ groups (i.e., the i^{th} level of that one factor), and $j = 1, 2,\ldots, n$ observations; and to test the multiple-level effect, H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$. The SAS procedure PROC ANOVA can be used for such multiple-sample test.

When the response variable (or outcome measure) holds the assumptions of independent, normally distributed with homogeneous (equal variance), then the One-Way ANOVA method can be implemented via the SAS procedure, PROC ANOVA with means statement, using the option /hovtest. See Zimmerman (2004), who discussed preliminary tests of equality of variances.

Similarly, when the response variable (or outcome measure) holds the assumptions of independent and normally distributed with non-homogeneous (or heterogeneous or unequal) variances, then the Welch (1947) ANOVA method can be implemented via the SAS procedure, PROC ANOVA with means statement, using the option /welch.

Wilcoxon (1945) and Mann and Whitney (1947) proposed a distribution-free model (i.e., nonparametric statistical methods) where the null hypothesis can be written as H_0: $F_1(X) = F_2(X)$ where $F_i(X)$ is the distribution function for sample $i = 1, 2$. This null hypothesis is to test whether the two population distributions are identical by using the sum of the ranks in sample 1 and sample 2. The test statistic is called the Wilcoxon rank-sum test (Mann-Whitney test). Alternatively, Zhao, Rahardja, and Qu (2008) considered quantifying the difference between the two groups, and defined the hypothesis in terms of the competing probability, $\pi = \Pr(X > Y) + 0.5 \Pr(X = Y)$, where X and Y are random variables with cumulative distribution functions (CDFs) F_X and F_Y, respectively. Then the following null hypothesis indicates there is no difference between the two groups: H_0: $\pi = 0.5$. Here the SAS procedure used is the PROC NPAR1WAY with Wilcoxon statement. For the distribution-free model (i.e., nonparametric statistical

methods) with multiple levels (multiple samples) within that one factor (or the grouping variable), the Kruskal-Wallis test of H_0: $F_1(X) = F_2(X) = \ldots = F_k(X)$ can be used (Kruskal & Wallis, 1952). Here the SAS PROC NPAR1WAY can be used.

Cao and Zhang (2014) reviewed various multiple comparison procedures (MCPs). Typically these MCPs are a part of an omnibus test (a series of sequential tests). For example, if using PROC GLM yields a statistically significant result for a main effect (or for an interaction, in the case of a two-factor or more scenarios), then one could use PROC MULTTEST to conduct the (pairwise) multiple comparisons. This PROC MULTTEST gives the raw *p*-values adjusted by Holm, Hochberg, or false discovery rate (FDR) methods. Note that under the LSMEANS statement of the PROC GLM, the "Adjust = BON;" option indicates the Bonferroni method. Among many of the above MCPs, the most commonly-used ones are Tukey's pairwise comparison, Bonferroni's method, Duncan, etc., depending on the specific needs, assumptions, or objective of the practitioners/researchers. For example, Tukey's method controls the Type I experiment-wise error rate and Bonferroni, Tukey's Least Significant Difference (LSD), and Duncan control the Type I comparison-wise error rate. Bonferroni has a very conservative (very wide) interval, i.e., is very slow to reject the null hypothesis. Table 1 summarizes the above discussion.

Roadmap

Provided in Figure 1 is the (two-sample and multiple-sample) roadmap for practitioners and researchers to choose a suitable testing method for their continuous (outcome measure or response variable) data analysis. In Figure 1, the roadmap method is provided by whether or not the response variable (outcome measure) is independent, then by whether or not the outcome is normally-distributed data, and then, finally, by whether or not the outcome variable has homogeneous variance. Then either yes/no response variable (in each of the 3 aforementioned questions) will lead to whether the grouping variable (or factor) is two-sample for a two-level factor or is multiple-sample or *k*-sample (where *k* is greater than 2) for a multiple-level factor. Next, the corresponding SAS procedures to the suitable statistical method directed from Figure 1 can be found in the Table 1 prescription.

Conclusion

Continuous data response or outcome is very common in real-data applications such as clinical trials, financial data, epidemiology, sociology, etc. The analysis of such continuous outcome measure (or response variable) has a long history, beginning with the one-sample *t*-test, two-sample *t*-test, up to the MCP. A review of the hypothesis testing procedures that are available for various types of continuous data outcome measure (or response variable) with one grouping variable (factor) of multiple levels are reviewed, along with the corresponding statistical computing translations/implementation in SAS, the most commonly used professional statistical software for data analysis.

Disclaimer

This research represents the author's own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other U.S. Federal agency.

References

Cao, J., & Zhang, S. (2014). Multiple comparison procedures. *The Journal of the American Medical Association, 312*(5), 543-544. doi: 10.1001/jama.2014.9440

Casella, G. (2008). *Statistical design*. New York, NY: Springer.

David, H. A., & Gunnink, J. L. (1997). The paired *t* test under artificial pairing. *The American Statistician 51*(1), 9-12. doi: 10.1080/00031305.1997.10473578

Kruskal, W., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association 47*(260), 583-621. doi: 10.2307/2280779

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*(1), 50-60. doi: 10.1214/aoms/1177730491

Rahardja, D., Yang, Y., & Zhang, Z. (2016). A comprehensive review of the two-sample independent or paired binary data – With or without stratum effects. *Journal of Modern Applied Statistical Methods, 15*(2), 215-223. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol15/iss2/16/

Welch, B. L. (1947). The generalization of ‘Student's’ problem when several different population variances are involved. *Biometrika, 34*(1/2), 28-35. doi: 10.2307/2332510

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83. doi: 10.2307/3001968

Zhao, Y. D., Rahardja, D., & Qu, Y. (2008). Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties. *Statistics in Medicine, 27*(3), 462-468. doi: 10.1002/sim.2912

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*(1), 173-181. doi: 10.1348/000711004849222

Appendix A: Tables and Figures

Table 1. Listing of response variable (outcome measure) type with the appropriate hypothesis testing, test statistic, and SAS command

Response (Outcome) Type/Assumptions	Null Hypothesis (H_0)	Test statistics	SAS command or other option
Independent, normal, homogeneous variance	$Y_{ij} = \mu_i + \epsilon_{ij}$ (cell means model) $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ (effect model) where $i = 1, 2,\ldots, k$ group, $j = 1, 2,\ldots, n_i$ observation		
Grouping variable: two-sample	H_0: $\mu_1 = \mu_2$ (cell means model) H_0: $\alpha_1 = \alpha_2$ (effect model)	Two-sample *t*-test (S-pooled)	PROC TTEST with class statement
Grouping variable: *k*-sample	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (cell means model) H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$ (effect model)	One-Way ANOVA	PROC ANOVA with means statement, using /hovtest option
Independent, normal, non-homogeneous variance	$Y_{ij} = \mu_i + \epsilon_{ij}$ (cell means model) $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ (effect model) where $i = 1, 2,\ldots, k$ group, $j = 1, 2,\ldots, n_i$ observation		
Grouping variable: two-sample	H_0: $\mu_1 = \mu_2$ (cell means model) H_0: $\alpha_1 = \alpha_2$ (effect model)	2-sample *t*-test (Satterthwaite exact d.f.)	PROC TTEST using /cochran option
Grouping variable: *k*-sample	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (cell means model) H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$ (effect model)	Welch ANOVA	PROC ANOVA using /welch option, under the means statement

Table 1, continued

Response (Outcome) Type/Assumptions	Null Hypothesis (H_0)	Test statistics	SAS command or other option
Independent, non-normal	Distribution shapes are the same but unspecified (distribution-free model)		
Grouping variable: 2-sample	2 Identical Distributions: H_0: $F_1(X) = F_2(X)$ Difference between 2 groups using competing probability: H_0: $\pi = 0.5$, where $\pi = P(X_1 > X_2) + 0.5\ P(X_1 = X_2)$ with random variables X_1, X_2 with CDFs F_1, F_2, respectively	Wilcoxon rank-sum test (Mann-Whitney test)	PROC NPAR1WAY with wilcoxon statement
Grouping variable: k-sample	H_0: $F_1(X) = F_2(X) = \ldots = F_k(X)$	Kruskal-Wallis Test	PROC NPAR1WAY
Not independent			
Grouping variable: two-sample	H_0: $\delta = 0$ $\delta = (\mu_1 - \mu_2)$	Paired t-test	PROC TTEST with paired statement
Grouping variable: k-sample	H_0: $\delta_1 = \delta_2 = \ldots = \delta_k$ where $\delta_i = (\mu_{i,1} - \mu_{i,2})$ $i = 1, \ldots, k$	Various MCPs such as Bonferroni, Tukey's LSD, Duncan, etc. See Cao and Zhang (2014)	Omnibus Test: PROC GLM using /Adjust=BON; option, under the LSMEANS statement PROC MULTTEST

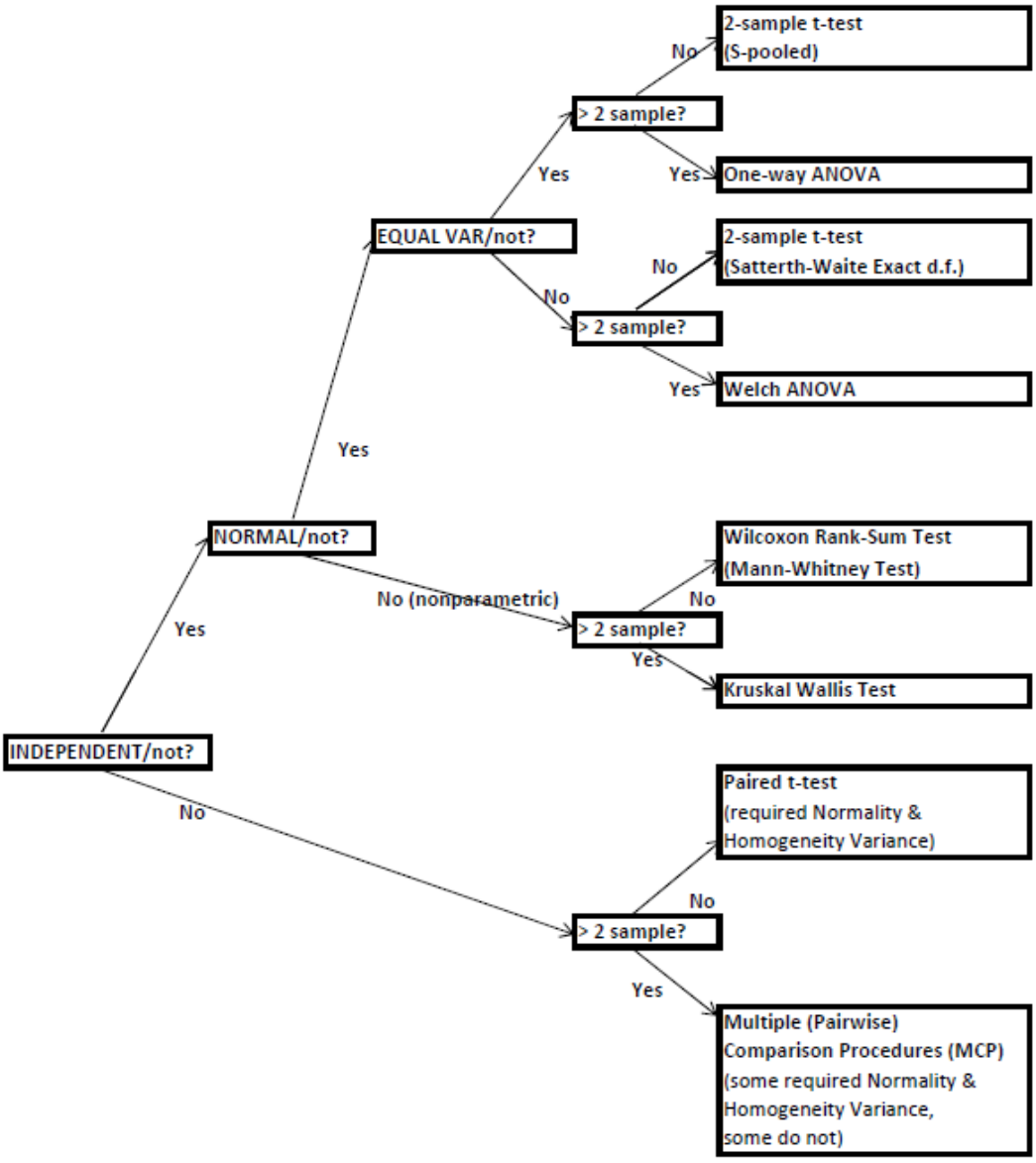


Figure 1. Continuous-data roadmap for two-sample and multiple-sample testing

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 137-157.
doi: 10.22237/jmasm/1493597280

ISSN 1538 − 9472

Test Statistics for the Comparison of Means for Two Samples That Include Both Paired and Independent Observations

Ben Derrick
University of the West Of England
Bristol, England, UK

Bethan Russ
Office for National Statistics
Newport, Wales, UK

Deirdre Toher
University of the West Of England
Bristol, England, UK

Paul White
University of the West Of England
Bristol, England, UK

Standard approaches for analyzing the difference in two means, where partially overlapping samples are present, are less than desirable. Here are introduced two test statistics, making reference to the *t*-distribution. It is shown that these test statistics are Type I error robust, and more powerful than standard tests.

Keywords: partially overlapping samples, test for equality of means, corrected *z*-test, partially correlated data, partially matched pairs

Introduction

Hypothesis tests for the comparison of two population means, μ_1 and μ_2, with two samples of either independent observations or paired observations are well established. When the assumptions of the test are met, the independent samples *t*-test is the most powerful test for comparing means between two independent samples (Sawilowsky and Blair, 1992). Similarly, when the assumptions of the test are met, the paired samples *t*-test is the most powerful test for the comparison of means between two dependent samples (Zimmerman, 1997). If a paired design

Ben Derrick is a Lecturer with the Applied Statistics Group. Email at ben.derrick@uwe.ac.uk. Bethan Russ is an associate with the Applied Statistics Group at UWE. Email at bethan.russ@ons.gsi.gov.uk. Dr. Toher is a Senior Lecturer with the Applied Statistics Group. Email at deirdre.toher@uwe.ac.uk. Dr. White is an Associate Professor and the academic lead for the Applied Statistics Group. Email at paul.white@uwe.ac.uk.

can avoid extraneous systematic bias, then paired designs are generally considered to be advantageous when contrasted with independent designs.

There are scenarios where, in a paired design, some observations may be missing. In the literature, this scenario is referred to as paired samples that are either "incomplete" (Ekbohm, 1976) or with "missing observations" (Bhoj, 1978). There are designs that do not have completely balanced pairings. Occasions where there may be two samples with both paired observations and independent observations include:

i) Two groups with some common element between both groups. For example, in education when comparing the average exam marks for two optional subjects, where some students take one of the two subjects and some students take both.

ii) Observations taken at two points in time, where the population membership changes over time but retains some common members. For example, an annual survey of employee satisfaction may include new employees that were unable to respond at time point one, employees that left after time point one, and employees that remained in employment throughout.

iii) When some natural pairing occurs. For example, in a survey taken comparing views of males and females, there will be some matched pairs (couples) and some independent individuals (single).

The examples given above can be seen as part of the wider missing data framework. There is much literature on methods for dealing with missing data and the proposals in this paper do not detract from extensive research into the area. The simulations and discussion in this paper are done in the context of data missing completely at random (MCAR).

Two samples that include both paired and independent observations is referred to using varied terminology in the literature. The example scenarios outlined can be referred to as "partially paired data" (Samawi and Vogel, 2011). However, this terminology has connotations suggesting that the pairs themselves are not directly matched. Derrick et al. (2015) suggest that appropriate terminology for the scenarios outlined gives reference to "partially overlapping samples." For work that has previously been done on a comparison of means when partially overlapping samples are present, "the partially overlapping

samples framework… has been treated poorly in the literature" (Martínez-Camblor, Corral, and María de la Hera, 2012, p.77). In this paper, the term partially overlapping samples will be used to refer to scenarios where there are two samples with both paired and independent observations.

When partially overlapping samples exist, the goal remains to test the null hypothesis $H_0 : \mu_1 = \mu_2$. Standard approaches when faced with such a situation, are to perform the paired samples *t*-test, discarding the unpaired data, or alternatively perform the independent samples *t*-test, discarding the paired data (Looney and Jones, 2003). These approaches are wasteful and can result in a loss of power. The bias created with these approaches may be of concern. Other solutions proposed in a similar context are to perform the independent samples *t*-test on all observations ignoring the fact that there may be some pairs, or alternatively randomly pairing unpaired observations and performing the paired samples *t*-test (Bedeian and Feild, 2002). These methods distort Type I error rates (Zumbo, 2002) and fail to adequately reflect the design. This emphasizes the need for research into a statistically valid approach. A method of analysis that takes into account any pairing but does not lose the unpaired information would be beneficial.

One analytical approach is to separately perform both the paired samples *t*-test on the paired observations and the independent samples *t*-test on the independent observations. The results are then combined using Fisher's (1925) Chi-square method, or Stouffer's (Stouffer, et al., 1949) weighted *z*-test. These methods have issues with respect to the interpretation of the results. Other procedures weighting the paired and independent samples *t*-tests, for the partially overlapping samples scenario, have been proposed by Bhoj, (1978), Kim et al. (2005), Martínez-Camblor, Corral, and María de la Hera (2012), and Samawi and Vogel (2011).

Looney and Jones (2003) proposed a statistic making reference to the *z*-distribution that uses all of the available data, without a complex weighting structure. Their corrected *z*-statistic is simple to compute and it directly tests the hypothesis $H_0 : \mu_1 = \mu_2$. They suggest that their test statistic is generally Type I error robust across the scenarios that they simulated. However, they only consider normally distributed data with a common variance of 1 and a total sample size of 50 observations. Therefore their simulation results are relatively limited, simulations across a wider range of parameters would help provide stronger conclusions. Mehrotra (2004) indicates that the solution provided by Looney and Jones (2003) may not be Type I error robust for small sample sizes.

Early literature for the partially overlapping samples framework focused on maximum likelihood estimates, when data are missing by accident rather than by design. Lin (1973) use maximum likelihood estimates for the specific case where data is missing from one of the two groups. Lin (1973) uses assumptions such as the variance ratio is known. Lin and Strivers (1974) apply maximum likelihood solutions to the more general case, but find that no single solution is applicable.

For normally distributed data, Ekbohm (1976) compared Lin and Strivers (1974) tests with similar proposals based on maximum likelihood estimators. Ekbohm (1976) found that maximum likelihood solutions do not always maintain Bradley's liberal Type I error robustness criteria. The results suggest that the maximum likelihood approaches are of little added value compared to standard methods. Furthermore the proposals by Ekbohm (1976) are complex mathematical procedures and are unlikely to be considered as a first choice solution in a practical environment.

A solution available in most standard software is to perform a mixed model using all of the available data. In a mixed model, effects are assessed using Restricted Maximum Likelihood estimators (REML). Mehrotra (2004) indicates that for positive correlation, REML is Type I error robust and more powerful approach than that proposed by Looney and Jones (2003).

For small sample sizes, an intuitive solution to the comparison of means with partially overlapping samples, would be a test statistic derived using concepts similar to that of Zumbo (2002) so that all available data are used making reference to the *t*-distribution.

Here, two test statistics are proposed. The proposed solution for equal variances acts as a linear interpolation between the paired samples *t*-test and the independent samples *t*-test. The consensus in the literature is that Welch's test is more Type I error robust than the independent samples *t*-test, particularly with unequal variances and unequal samples sizes (Derrick, Toher and White, 2016; Fay and Proschan, 2010; Zimmerman and Zumbo, 2009). The proposed solution for unequal variances is a test that acts as a linear interpolation between the paired samples *t*-test and Welch's test.

Standard tests and the proposal by Looney and Jones (2003) are given below. This is followed by the definition of the presently proposed test statistics. A worked example using each of these test statistics and REML is provided. The Type I error rate and power for the test statistics and REML is then explored using simulation, for partially overlapping samples simulated from a Normal distribution.

Notation

Notation used in the definition of the test statistics is given in Table 1.

Table 1. Notation used in this paper.

n_a = number of observations exclusive to Sample 1
n_b = number of observations exclusive to Sample 2
n_c = number of pairs
n_1 = total number of observations in Sample 1 (i.e. $n_1 = n_a + n_c$)
n_2 = total number of observations in Sample 2 (i.e. $n_2 = n_b + n_c$)
$\bar{X}_1$ = mean of all observations in Sample 1
$\bar{X}_2$ = mean of all observations in Sample 2
$\bar{X}_a$ = mean of the independent observations in Sample 1
$\bar{X}_b$ = mean of the independent observations in Sample 2
$\bar{X}_{1c}$ = mean of the paired observations in Sample 1
$\bar{X}_{2c}$ = mean of the paired observations in Sample 2
S_1^2 = variance of all observations in Sample 1
S_2^2 = variance of all observations in Sample 2
S_a^2 = variance of the independent observations in Sample 1
S_b^2 = variance of the independent observations in Sample 2
S_{1c}^2 = variance of the paired observations in Sample 1
S_{2c}^2 = variance of the paired observations in Sample 2
S_{12} = covariance between the paired observations
r = Pearson's correlation coefficient for the paired observations

All variances above are calculated using Bessel's correction, i.e. the sample variance with $n_i - 1$ degrees of freedom (see Kenney and Keeping, 1951, p.161).

As standard notation, random variables are shown in upper case, and derived sample values are shown are in lower case.

Definition of Existing Test Statistics

Standard approaches for comparing two means making reference to the *t*-distribution are given below. These definitions follow the structural form given by Fradette et al. (2003), adapted to the context of partially overlapping samples.

To perform the paired samples *t*-test, the independent observations are discarded so that

$$T_1 = \frac{\bar{X}_{1c} - \bar{X}_{2c}}{\sqrt{\frac{S_{1c}^2}{n_c} + \frac{S_{2c}^2}{n_c} - 2r\left(\frac{S_{1c}S_{2c}}{n_c}\right)}}$$

The statistic T_1 is referenced against the *t*-distribution with $\nu_1 = n_c - 1$ degrees of freedom.

To perform the independent samples *t*-test, the paired observations are discarded so that

$$T_2 = \frac{\bar{X}_a - \bar{X}_b}{S_p\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \text{ where } S_p = \sqrt{\frac{(n_a - 1)S_a^2 + (n_b - 1)S_b^2}{(n_a - 1) + (n_b - 1)}}$$

The statistic T_2 is referenced against the *t*-distribution with $\nu_2 = n_a + n_b - 2$ degrees of freedom.

To perform Welch's test, the paired observations are discarded so that

$$T_3 = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}}}$$

The statistic T_3 is referenced against the *t*-distribution with degrees of freedom approximated by

$$v_3 = \frac{\left(\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}\right)^2}{\left(\frac{S_a^2}{n_a}\right)^2 / (n_a - 1) + \left(\frac{S_b^2}{n_b}\right)^2 / (n_b - 1)}$$

For large sample sizes, the test statistic for partially overlapping samples proposed by Looney and Jones (2003) is

$$Z_{\text{corrected}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_a + n_c} + \frac{S_2^2}{n_b + n_c} - \frac{(2n_c)S_{12}}{(n_a + n_c)(n_b + n_c)}}}$$

The statistic $Z_{\text{corrected}}$ is referenced against the standard Normal distribution. In the extremes of $n_a = n_b = 0$, or $n_c = 0$, $Z_{\text{corrected}}$ defaults to the paired samples z-statistic and the independent samples z-statistic respectively.

Definition of Proposed Test Statistics

Two new t-statistics are proposed; T_{new1}, assuming equal variances, and T_{new2}, when equal variances cannot be assumed. The test statistics are constructed as the difference between two means taking into account the covariance structure. The numerator is the difference between the means of the two samples and the denominator is a measure of the standard error of this difference. Thus the test statistics proposed here are directly testing the hypothesis $H_0 : \mu_1 = \mu_2$.

The test statistic T_{new1} is derived so that in the extremes of $n_a = n_b = 0$, or $n_c = 0$, T_{new1} defaults to T_1 or T_2 respectively, thus

$$T_{\text{new1}} = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\left(\frac{n_c}{n_1 n_2}\right)}} \text{ where } S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

The test statistic T_{new1} is referenced against the t-distribution with degrees of freedom derived by linear interpolation between v_1 and v_2 so that

$$v_{\text{new1}} = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c} \right)(n_a + n_b).$$

In the extremes, when $n_a = n_b = 0$, v_{new1} defaults to v_1; or when $n_c = 0$, v_{new1} defaults to v_2.

Given the superior Type I error robustness of Welch's test when variances are not equal, a test statistic is derived making reference to Welch's approximate degrees of freedom. This test statistic makes use of the sample variances, S_1^2 and S_2^2. The test statistic T_{new2} is derived so that in the extremes of $n_a = n_b = 0$, or $n_c = 0$, T_{new2} defaults to T_1 or T_3 respectively, thus

$$T_{\text{new2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2} \right)}}$$

The test statistic T_{new2} is referenced against the *t*-distribution with degrees of freedom derived as a linear interpolation between v_1 and v_3 so that

$$v_{\text{new2}} = (n_c - 1) + \left(\frac{\gamma - n_c - 1}{n_a + n_b + 2n_c} \right)(n_a + n_b)$$

$$\text{where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left(\frac{S_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{S_2^2}{n_2} \right)^2 / (n_2 - 1)}$$

In the extremes, when $n_a = n_b = 0$, v_{new2} defaults to v_1; or when $n_c = 0$, v_{new2} defaults to v_3.

Note that the proposed statistics, T_{new1} and T_{new2}, use all available observations in the respective variance calculations. The statistic $Z_{\text{corrected}}$ only uses the paired observations in the calculation of covariance.

Worked Example

An applied example is given to demonstrate the calculation of each of the test statistics defined. In education, for credit towards an undergraduate Statistics course, students may take optional modules in either Mathematical Statistics, or Operational Research, or both. The program leader is interested whether the exam marks for the two optional modules differ. The exam marks attained for a single semester are given in Table 2.

Table 2. Exam marks for students studying on an undergraduate Statistics course.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Mathematical Statistics	73	82	74	59	49	-	42	71	-	39	-	-	-	-	59	85
Operational Research	72	-	89	78	64	83	42	76	79	89	67	82	85	92	63	-

As per standard notion, the derived sample values are given in lower case. In the calculation of the test statistics, $\bar{x}_1 = 63.300$, $\bar{x}_2 = 75.786$, $s_1^2 = 263.789$, $s_2^2 = 179.874$, $n_a = 2$, $n_b = 6$, $n_c = 8$, $n_1 = 10$, $n_2 = 14$, $v_1 = 7$, $v_2 = 6$, $v_3 = 6$, $\gamma = 17.095$, $v_{\text{new1}} = 12$, $v_{\text{new2}} = 10.365$, $r = 0.366$, $s_{12} = 78.679$.

For the REML analysis, a mixed model is performed with "Module" as a repeated measures fixed effect and "Student" as a random effect. Table 3 gives the calculated test statistics, degrees of freedom and corresponding *p*-values.

Table 3. Test statistic values and resulting *p*-values (two-sided test).

	T_1	T_2	T_3	$Z_{\text{corrected}}$	REML	T_{new1}	T_{new2}
estimate of mean difference	-13.375	2.167	2.167	-12.486	-12.517	-12.486	-12.486
***t*-value**	-2.283	0.350	0.582	-2.271	-2.520	-2.370	-2.276
degrees of freedom	7.000	6.000	6.000		11.765	12.000	10.365
***p*-value**	0.056	0.739	0.579	0.023	0.027	0.035	0.045

With the exception of REML, the estimates of the mean difference are simply the difference in the means of the two samples, based on the observations used in the calculation. It can quickly be seen that the conclusions differ depending on the test used. It is of note that only the tests using all of the available data result in the rejection of the null hypothesis at $\alpha_{\text{nominal}} = 0.05$. Also note that the results of the paired samples *t*-test and the independent samples *t*-test have sample effects in different directions. This is only one specific example

given for illustrative purposes, investigation is required into the power of the test statistics over a wide range of scenarios. Conclusions based on the proposed tests cannot be made without a thorough investigation into their Type I error robustness.

Simulation Design

Under normality, Monte-Carlo methods are used to investigate the Type I error robustness of the defined test statistics and REML. Power should only be used to compare tests when their Type I error rates are equal (Zimmerman and Zumbo, 1993). Monte-Carlo methods are used to explore the power for the tests that are Type I error robust under normality.

Unbalanced designs are frequent in psychology (Sawilowsky and Hillman, 1992), thus a comprehensive range of values for n_a, n_b and n_c are simulated. These values offer an extension to the work done by Looney and Jones (2003). Given the identification of separate test statistics for equal and unequal variances, multiple population variance parameters $\{\sigma_1^2, \sigma_2^2\}$ are considered. Correlation has an impact on Type I error and power for the paired samples *t*-test (Fradette et al., 2003), hence a range of correlations $\{\rho\}$ between two normal populations are considered. Correlated normal variates are obtained as per Kenney and Keeping (1951). A total of 10,000 replicates of each of the scenarios in Table 4 are performed in a factorial design.

All simulations are performed in R version 3.1.2. For the mixed model approach utilizing REML, the R package `lme4` is used. Corresponding p-values are calculated using the R package `lmerTest`, which uses the Satterthwaite approximation adopted by SAS (Goodnight, 1976).

For each set of 10,000 *p*-values, the proportion of times the null hypothesis is rejected, for a two sided test with $\alpha_{\text{nominal}} = 0.05$ is calculated.

Table 4. Summary of simulation parameters

Parameter	Values
μ_1	0
μ_2	0 (under H_0); 0.5 (under H_1)
σ_1^2	1, 2, 4, 8
σ_2^2	1, 2, 4, 8
n_a	5, 10, 30, 50, 100, 500
n_b	5, 10, 30, 50, 100, 500
n_c	5, 10, 30, 50, 100, 500
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75

Type I Error Robustness

For each of the test statistics, Type I error robustness is assessed against Bradley's (1978) liberal criteria. This criteria is widely used in many studies analyzing the validity of *t*-tests and their adaptions. Bradley's (1978) liberal criteria states that the Type I error rate α should be within $\alpha_{\text{nominal}} \pm 0.5\,\alpha_{\text{nominal}}$. For $\alpha_{\text{nominal}} = 0.05$, Bradley's liberal interval is [0.025, 0.075].

Type I error robustness is firstly assessed under the condition of equal variances. Under the null hypothesis, 10,000 replicates are obtained for the $4 \times 6 \times 6 \times 6 \times 7 = 6{,}048$ scenarios where $\sigma_1^2 = \sigma_2^2$. Figure 1 shows the Type I error rates for each of the test statistics under equal variances for normally distributed data.

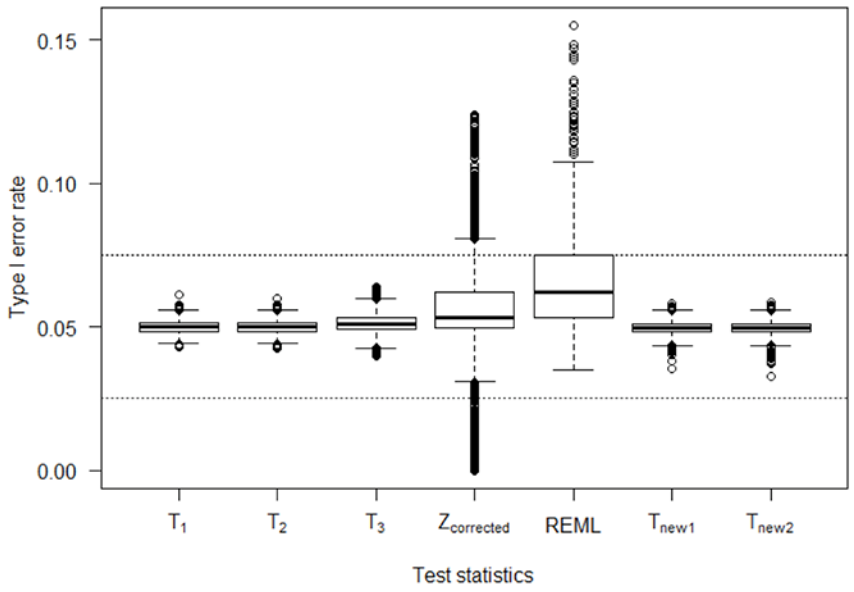


Figure 1. Type I error rates where $\sigma_1^2 = \sigma_2^2$, reference lines show Bradley's (1978) liberal criteria.

Figure 1 indicates that when variances are equal, the statistics T_1, T_2, T_3, T_{new1} and T_{new2} remain within Bradley's liberal Type I error robustness criteria throughout the entire simulation design. The statistic $Z_{\text{corrected}}$ is not Type I error robust, thus confirming the smaller simulation findings of Mehrotra (2004). Figure 1 also shows that REML is not Type I error robust throughout the entire

simulation design. A review of our results shows that for REML the scenarios that are outside the range of liberal Type I error robustness are predominantly those that have negative correlation, and some where zero correlation is specified. Given that negative correlation is rare in a practical environment, the REML procedure is not necessarily unjustified.

Type I error robustness is assessed under the condition of unequal variances. Under the null hypothesis, 10,000 replicates were obtained for the $4 \times 3 \times 6 \times 6 \times 6 \times 7 = 18{,}144$ scenarios where $\sigma_1^2 \neq \sigma_2^2$. For assessment against Bradley's (1978) liberal criteria, Figure 2 shows the Type I error rates for unequal variances for normally distributed data.

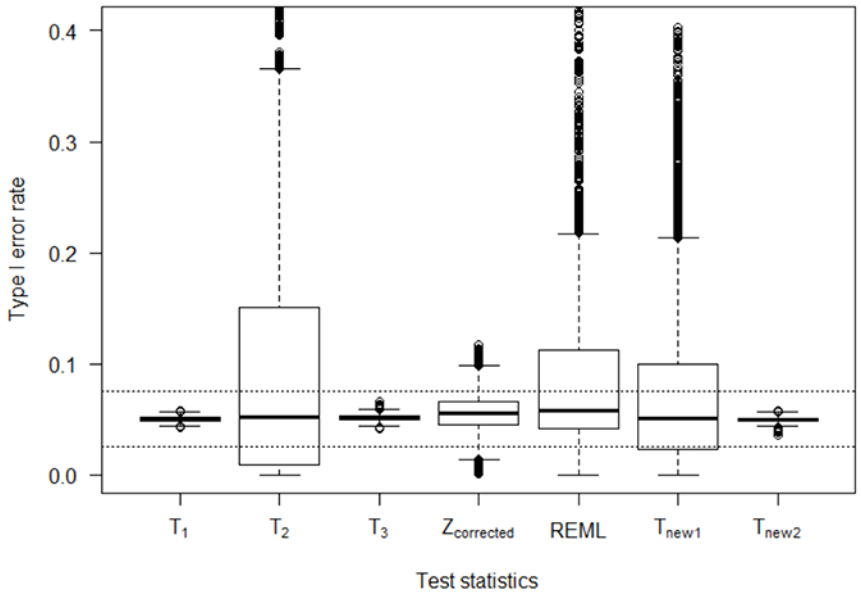


Figure 2. Type I error rates when $\sigma_1^2 \neq \sigma_2^2$, reference lines show Bradley's (1978) liberal criteria.

Figure 2 illustrates that that the statistics defined using a pooled standard deviation, T_2 and T_{new1}, do not provide Type I error robust solutions when equal variances cannot be assumed. The statistics T_1, T_3 and T_{new2} retain their Type I error robustness under unequal variances throughout all conditions simulated.

The statistic $Z_{corrected}$ maintains similar Type I error rates under equal and unequal variances. The statistic $Z_{corrected}$ was designed to be used only in the case

of equal variances. For unequal variances, we observe that the statistic $Z_{\text{corrected}}$ results in an unacceptable amount of false positives when $\rho \leq 0.25$ or $\max\{n_a, n_b, n_c\} - \min\{n_a, n_b, n_c\}$ is large. In addition, the statistic $Z_{\text{corrected}}$ is conservative when ρ is large and positive. The largest observed deviations from Type I error robustness for REML are when $\rho \leq 0$ or $\max\{n_a, n_b, n_c\} - \min\{n_a, n_b, n_c\}$ is large. Further insight to the Type I error rates for REML can be seen in Figure 3 showing observed *p*-values against expected *p*-values from a uniform distribution.

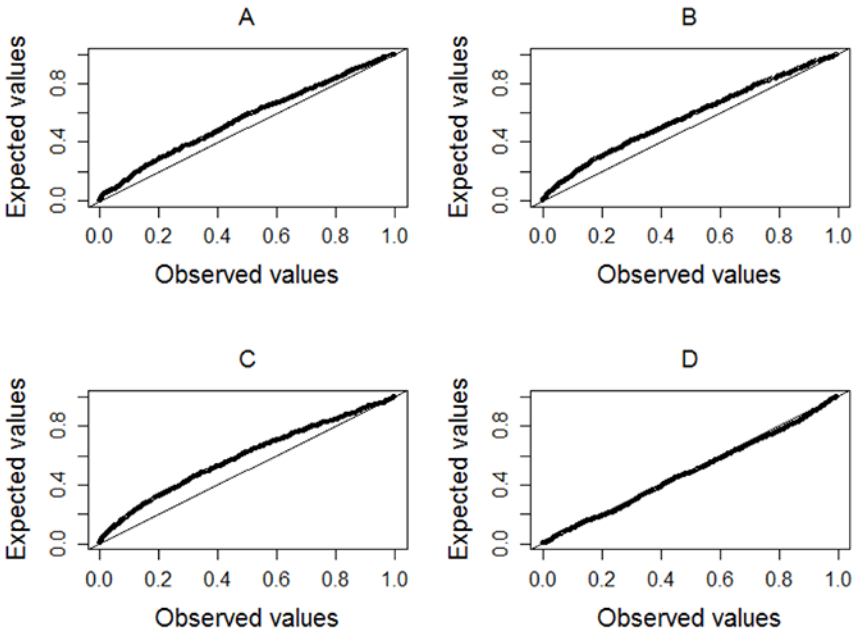


Figure 3. P-P plots for simulated *p*-values using REML procedure. Selected parameter combinations (n_a, n_b, n_c, σ_1^2, σ_2^2, ρ) are as follows; *A* = (5,5,5,1,1,-0.75), *B* = (5,10,5,8,1,0), *C* = (5,10,5,8,1,0.5), *D* = (10,5,5,8,1,0.5).

If the null hypothesis is true, for any given set of parameters the *p*-values should be uniformly distributed. Figure 3 gives indicative parameter combinations where the *p*-values are not uniformly distributed when applying a mixed model assessed using REML. It can be seen that REML is not Type I error robust when the correlation is negative. In addition, caution should be exercised if using REML when the larger variance is associated with the smaller sample size.

REML maintains Type I error robustness for positive correlation and equal variances or when the larger sample size is associated with the larger variance.

Power of Type I Error Robust Tests under Equal Variances

The test statistics that do not fail to maintain Bradley's Type I error liberal robustness criteria are assessed under H_1. REML is included in the comparisons for $\rho \geq 0$. The power of the test statistics are assessed where $\sigma_1^2 = \sigma_2^2 = 1$, followed by an assessment of the power of the test statistics where $\sigma_1^2 > 1$ and $\sigma_2^2 = 1$.

Table 5 shows the power of T_1, T_2, T_3, T_{new1}, T_{new2} and REML, averaged over all sample size combinations where $\sigma_1^2 = \sigma_2^2 = 1$.

Table 5. Power of Type I error robust test statistics $\sigma_1^2 = \sigma_2^2 = 1$, $\alpha = 0.05$, $\mu_2 - \mu_1 = 0.5$.

	ρ	T_1	T_2	T_3	T_{new1}	T_{new2}	REML
	0.75	0.785	0.567	0.565	0.887	0.886	0.922
	0.50	0.687	0.567	0.565	0.865	0.864	0.880
$n_a = n_b$	0.25	0.614	0.567	0.565	0.842	0.841	0.851
	0	0.558	0.567	0.565	0.818	0.818	0.829
	<0	0.481	0.567	0.565	0.778	0.778	-
	0.75	0.784	0.455	0.433	0.855	0.847	0.907
	0.5	0.687	0.455	0.433	0.84	0.832	0.861
$n_a \neq n_b$	0.25	0.615	0.455	0.433	0.823	0.816	0.832
	0	0.559	0.455	0.433	0.806	0.799	0.816
	<0	0.482	0.455	0.433	0.774	0.766	-

Table 5 shows that REML and the test statistics proposed in this paper, T_{new1} and T_{new2}, are more powerful than standard approaches, T_1, T_2 and T_3, when variances are equal. Consistent with the paired samples *t*-test, T_1, the power of T_{new1} and T_{new2} is relatively lower when there is zero or negative correlation between the two populations. Similar to contrasts of the independent samples *t*-test, T_2, with Welch's test, T_3, for equal variances but unequal sample sizes, T_{new1} is marginally more powerful than T_{new2}, but not to any practical extent. For each of the tests statistics making use of paired data, as the correlation between the paired samples increases, the power increases.

As the correlation between the paired samples increases, the power advantage of the proposed test statistics relative to the paired samples *t*-test becomes smaller. Therefore the proposed statistics T_{new1} and T_{new2} may be especially useful when the correlation between the two populations is small.

To show the relative increase in power for varying sample sizes, Figure 4 shows the power for selected test statistics for small-medium sample sizes, averaged across the simulation design for equal variances.

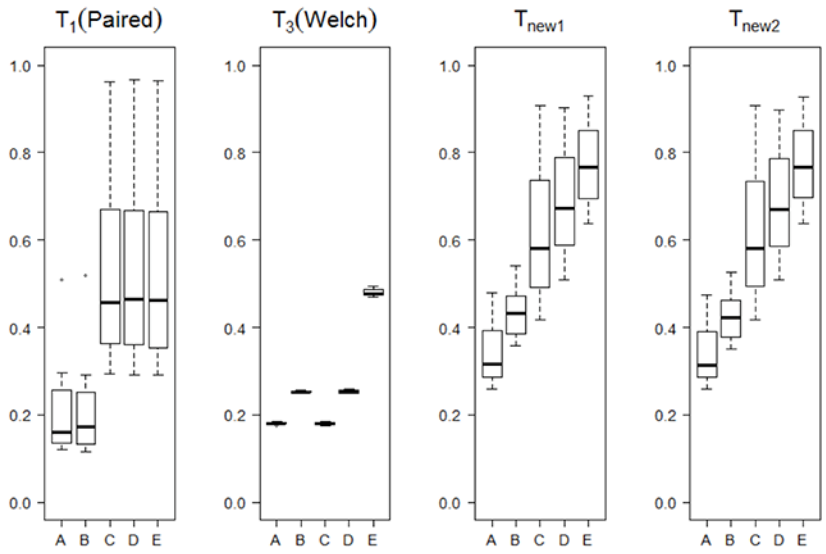


Figure 4. Power for Type I error robust test statistics, averaged across all values of ρ where $\sigma_1^2 = \sigma_2^2$ and $\mu_2 - \mu_1 = 0.5$. The sample sizes (n_a, n_b, n_c) are as follows: *A* = (10,10,10), *B* = (10,30,10), *C* = (10,10,30), *D* = (10,30,30), *E* = (30,30,30).

From Figure 4 it can be seen that for small-medium sample sizes, the power of the proposed test statistics T_{new1} and T_{new2} is superior to standard test statistics.

Power of Type I Error Robust Rests under Unequal Variances

For the Type I error robust test statistics under unequal variances, Table 6 describes the power of T_1, T_3, T_{new2} and REML, averaged over the simulation design where $\mu_2 - \mu_1 = 0.5$. Table 6 shows that T_{new2} has superior power properties to both T_1 and T_3 when variances are not equal. In common with the performance of Welch's test for independent samples, T_3, the power of T_{new2} is higher when the larger variance is associated with the larger sample size. In common with the performance of the paired samples *t*-test, T_1, the power of T_{new2} is relatively lower when there is zero or negative correlation between the two populations.

Table 6. Power of Type I error robust test statistics where $\sigma_1^2 > 1$, $\sigma_2^2 = 1$, $\alpha = 0.05$, $\mu_2 - \mu_1 = 0.5$. Within this table, $n_a > n_b$ represents the larger variance associated with the larger sample size, and $n_a < n_b$ represents the larger variance associated with the smaller sample size.

	ρ	T_1	T_3	T_{new2}	REML
$n_a = n_b$	0.75	0.555	0.393	0.692	0.645
	0.50	0.481	0.393	0.665	0.588
	0.25	0.429	0.393	0.640	0.545
	0	0.391	0.393	0.619	0.515
	<0	0.341	0.393	0.582	-
$n_a > n_b$	0.75	0.555	0.351	0.715	0.589
	0.50	0.481	0.351	0.688	0.508
	0.25	0.429	0.351	0.665	0.459
	0	0.391	0.351	0.642	0.422
	<0	0.341	0.351	0.604	-
$n_a < n_b$	0.75	0.555	0.213	0.559	0.693
	0.50	0.481	0.213	0.539	0.649
	0.25	0.429	0.213	0.522	0.62
	0	0.391	0.213	0.507	0.603
	<0	0.341	0.213	0.480	-

The apparent power gain for REML when the larger variance is associated with the larger sample size, can be explained by the pattern in the Type I error rates. REML follows a similar pattern to the independent samples *t*-test, which is liberal when the larger variance is associated with the larger sample size, thus giving the perception of higher power.

To show the relative increase in power for varying sample sizes, Figure 5 shows the power for selected test statistics for small-medium sample sizes, averaged across the simulation design for unequal variances.

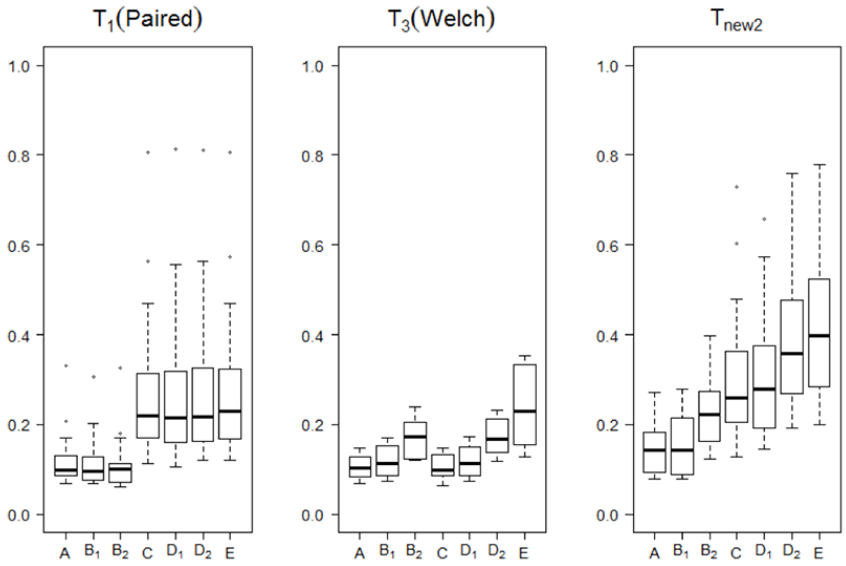


Figure 5. Power for Type I error robust test statistics $\sigma_1^2 > \sigma_2^2$ and $\mu_2 - \mu_1 = 0.5$. The sample sizes (n_a, n_b, n_c) are as follows: A = (10,10,10), B_1 = (10,30,10), B_2 = (30,10,10), C = (10,10,30), D_1 = (10,30,30), D_2 = (30,10,30), E = (30,30,30).

Figure 5 shows a relative power advantage when the larger variance is associated with the larger sample size, as per B_2 and D_2. A comparison of Figure 4 and Figure 5 shows that for small-medium sample sizes, power is adversely affected for all test statistics when variances are not equal.

Discussion

The statistic T_{new2} is Type I error robust across all conditions simulated under normality. The greater power observed for T_{new1}, compared to T_{new2}, under equal variances, is likely to be of negligible consequence in a practical environment. This is in line with empirical evidence for the performance of Welch's test, when only independent samples are present, which leads to many observers recommending the routine use of Welch's test under normality (e.g. Ruxton, 2006).

The Type I error rates and power of T_{new2} follow the properties of its counterparts, T_1 and T_3. Thus T_{new2} can be seen as a trade-off between the paired samples *t*-test and Welch's test, with the advantage of increased power across all conditions, due to using all available data.

The partially overlapping samples scenarios identified in this paper could be considered as part of the missing data framework and all simulations have been performed under the assumption of MCAR.

The statistics proposed in this paper form less computationally intensive competitors to REML. The REML procedure does not directly calculate the difference between the two sample means, in a practical environment this makes its results hard to interpret. The statistics proposed in this paper also lend themselves far more easily to the development of non-parametric tests.

Conclusion

A commonly occurring scenario when comparing two means is a combination of paired observations and independent observations in both samples, this scenario is referred to as partially overlapping samples. Standard procedures for analyzing partially overlapping samples involve discarding observations and performing either the paired samples *t*-test, or the independent samples *t*-test, or Welch's test. These approaches are less than desirable. In this paper, two new test statistics making reference to the *t*-distribution are introduced and explored under a comprehensive set of parameters, for normally distributed data. Under equal variances, T_{new1} and T_{new2} are Type I error robust. In addition they are more powerful than standard Type I error robust approaches considered in this paper. When variances are equal, there is a slight power advantage of using T_{new1} over T_{new2}, particularly when sample sizes are not equal. Under unequal variances, T_{new2} is the most powerful Type I error robust statistic considered in this paper. We recommend that when faced with a research problem involving partially overlapping samples and MCAR can be reasonably assumed, the statistic T_{new1} could be used when it is known that variances are equal. Otherwise under the same conditions when equal variances cannot be assumed the statistic T_{new2} could be used.

A mixed model procedure using REML is not fully Type I error robust. In those scenarios in which this procedure is Type I error robust, the power is similar to that of T_{new1} and T_{new2}.

The proposed test statistics for partially overlapping samples provide a real alternative method for analysis for normally distributed data, which could also be used for the formation of confidence intervals for the true difference in two means.

References

Bedeian, A. G., & Feild, H. S. (2002). Assessing group change under conditions of anonymity and overlapping samples. *Nursing research, 51*(1), 63-65. doi: 10.1097/00006199-200201000-00010

Bhoj, D. (1978). Testing equality of means of correlated variates with missing observations on both responses. *Biometrika, 65*(1), 225-228. doi: 10.1093/biomet/65.1.225

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Derrick, B., Dobson-McKittrick, A., Toher, D., & White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods, 10*(3)

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology, 12*(1), 30-38. doi: 10.20982/tqmp.12.1.p030

Ekbohm, G. (1976). On comparing means in the paired case with incomplete data on both responses, *Biometrika, 63*(2), 299-304. doi: 10.1093/biomet/63.2.299

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys, 4*(1). doi: 10.1214/09-SS051

Fisher, R. A. (1925). *Statistical methods for research workers*. New Delhi, India: Genesis Publishing Pvt. Ltd.

Fradette, K., Keselman, H. J., Lix, L., Algina, J., & Wilcox, R. (2003). Conventional and robust paired and independent samples *t*-tests: Type I error and power rates. *Journal of Modern Applied Statistical Methods, 2*(2), 481-496. doi: 10.22237/jmasm/1067646120

Kenney, J. F., & Keeping, E. S. (1951). *Mathematics of statistics, Pt. 2* (2nd Ed). Princeton, NJ: Van Nostrand.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., & Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics, 21*(4), 517-528. doi: 10.1093/bioinformatics/bti029

Lin, P. E. (1973). Procedures for testing the difference of means with incomplete data. *Journal of the American Statistical Association, 68*(343), 699-703. doi: 10.1080/01621459.1973.10481407

Lin, P. E., & Strivers L. (1974). Difference of means with incomplete data. *Biometrika, 61*(2), 325-334. doi: 10.1093/biomet/61.2.325

Looney, S., & Jones, P. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine, 22*, 1601-1610. doi: 10.1002/sim.1514

Martínez-Camblor, P., Corral, N., & María de la Hera, J. (2013). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics, 40*(1), 76-87. doi: 10.1080/02664763.2012.734795

Mehrotra, D. (2004). Letter to the editor, a method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine, 23*(7), 1179–1180. doi: 10.1002/sim.1693

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. 2014; version 3.1.2.

Ruxton, G. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann-Whitney U test. *Behavioral Ecology, 17*(4), 688. doi: 10.1093/beheco/ark016

Goodnight, J. H. (1976). *General linear models procedure*. S.A.S. Institute. Inc.

Samawi, H. M., & Vogel, R. (2011). Tests of homogeneity for partially matched-pairs data. *Statistical Methodology, 8*(3), 304-313. doi: 10.1016/j.stamet.2011.01.002

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the *t*-test to departures from population normality. *Psychological Bulletin, 111*(2), 352. doi: 10.1037/0033-2909.111.2.352

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples *t*-test under a prevalent psychometric measure distribution, Journal of *Consulting and Clinical Psychology, 60*(2), 240-243. doi: 10.1037/0022-006X.60.2.240

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). *The American soldier: adjustment during army life* (*Studies in*

Social Psychology in World War II, Vol. I). Princeton, NJ: Princeton University Press.

Zimmerman, D. W. (1997). A note on the interpretation of the paired samples *t*-test. *Journal of Educational and Behavioral Statistics, 22*(3), 349 – 360. doi: 10.3102/10769986022003349

Zimmerman, D. W., & Zumbo, B. D. (1993). Significance testing of correlation using scores, ranks, and modified ranks. *Educational and Psychological Measurement, 53*(4), 897-904. doi: 10.1177/0013164493053004003

Zimmerman, D. W., & Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variances t tests. *Psicológica: Revista de Metodología y Psicología Experimental, 30*(2), 371-390.

Zumbo, B. D. (2002). An adaptive inference strategy: The case of auditory data. *Journal of Modern Applied Statistical Methods, 1*(1), 60-68. doi: 10.22237/jmasm/1020255000

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 158-178.
doi: 10.22237/jmasm/1493597340

ISSN 1538 − 9472

Effective Estimation Strategy of Finite Population Variance Using Multi-Auxiliary Variables in Double Sampling

Reba Maji
Sarojini Naidu College for Women
Kolkata, India

G. N. Singh
Indian School of Mines
Dhanbad, India

Arnab Bandyopadhyay
Asansol Engineering College
Asansol, India

Estimation of population variance in two-phase (double) sampling is considered using information on multiple auxiliary variables. An unbiased estimator is proposed and its properties are studied under two different structures. The superiority of the suggested estimator over some contemporary estimators of population variance was established through empirical studies from a natural and an artificially generated dataset.

Keywords: Double sampling, study variable, auxiliary variable, chain-type, regression, bias, variance, efficiency

Introduction

Auxiliary information plays a role in the planning, selection, and estimation stages of a sample survey. Sometimes information on several auxiliary variables may be readily available. For instance, to study the case of public health and welfare of a state or a country, the number of beds in different hospitals, doctors, and supporting staffs may be known, as well as the amount of funds available for medicine. When such information is lacking, it may be possible to obtain a large preliminary sample in which the auxiliary variable is measured, which is the premise of two-phase sampling, also known as double sampling. It is a powerful and cost-effective technique for obtaining reliable estimates in the first phase sample for the unknown parameters of the auxiliary variables.

Variation is an inherent phenomenon of nature. The use of auxiliary information in the estimation of population variance was considered by Das and Tripathi (1978), and extended by Isaki (1983), R. K. Singh (1983), Srivastava and

Reba Maji is an Assistant Professor of Mathematics. Email them at: rebamaji09@gmail.com. G. N. Singh is in the Department of Applied Mathematics. Email them at: gnsingh_ism@yahoo.com. Arnab Bandyopadhyay is in the Department of Mathematics. Email them at: arnabbandyopadhyay4@gmail.com.

Jhaji (1980), Upadhyaya and Singh (1983), Tripathi, Singh, and Upadhyaya (1988), Prasad and Singh (1990, 1992), S. Singh and Joarder (1998), R. Singh, Chauhan, Sawan, and Smarandache (2011), and Tailor and Sharma (2012), among others. However, most of these estimators of population variance are biased and based on the assumptions that the population mean or variance of the auxiliary variables are known, which may become a serious drawback in estimating population parameters in sample surveys.

Motivated with the above arguments, the objective of the present work is to propose an efficient and unbiased estimator of the population variance. The properties of the proposed estimator have been studied under two different structures of double sampling and results are supported with suitable simulation studies carried over six real datasets and an artificially generated data set.

Formulation of the Proposed Estimator

Consider a finite population $\mathrm{U}=\left(U_1,U_2,\ldots,U_N\right)$. Let y be the character under study and x_i, $i = 1, 2,\ldots, p$, be p (non-negative integer constant) auxiliary variables, taking values y_h and x_{i_h}, respectively, for the h^{th} unit. We define

$$S_y^2=\frac{N}{N-1}\sigma_y^2 \text{ with } \sigma_y^2=\frac{1}{N}\sum_{h=1}^{N}\left(y_h-\bar{Y}\right)^2$$

$$S_{x_i}^2=\frac{N}{N-1}\sigma_{x_i}^2 \text{ with } \sigma_{x_i}^2=\frac{1}{N}\sum_{h=1}^{N}\left(x_{i_h}-\bar{X}_i\right)^2$$

where

$$\bar{Y}=\frac{1}{N}\sum_{h=1}^{N}y_h \text{ and } \bar{X}_i=\frac{1}{N}\sum_{h=1}^{N}x_{i_h},\ i=1,2,\ldots,p$$

are the population means of y and x_i, respectively. For large N, $S_y^2\cong\sigma_y^2$ and $S_{x_i}^2\cong\sigma_{x_i}^2\ \forall i = 1, 2,\ldots, p$.

Estimate the population variance S_y^2 of y when the population variances $S_{x_i}^2$ of x_i ($i = 1, 2,\ldots, p$) are unknown. When the variables y and the x_i are closely related but no information is available on the population variances $S_{x_i}^2$ of x_i, we seek to estimate S_y^2 from a sample S, obtained through a two-phase (or double)

selection. In this sampling scheme, a first phase sample S' (S' ⊂ U) of size n' is drawn by a simple random sampling without replacement (SRSWOR) scheme from the entire population U and the auxiliary variables x_i are observed to furnish the estimates of $S_{x_i}^2$ ($i = 1, 2, \ldots, p$). A second phase sample S of size n ($n \leq n'$) is drawn according to one of the following rules by the method of SRSWOR to observe the study variable y:

Case I: The second phase sample is drawn as a subsample of the first phase sample (i.e. S ⊂ S').

Case II: The second phase sample is drawn independently of the first phase sample.

Using one auxiliary variable x, Isaki (1983) suggested a ratio estimator for S_y^2 whose two-phase sampling version may be defined as

$$t_1 = s_y^2(n)\frac{s_x^2(n')}{s_x^2(n)} \tag{1}$$

where

$$s_y^2(n) = \frac{1}{n-1}\sum_{h=1}^{n}(y_h - \bar{y})^2,\ \bar{y} = \frac{1}{n}\sum_{h=1}^{n} y_h$$

$$s_x^2(n) = \frac{1}{n-1}\sum_{h=1}^{n}(x_h - \bar{x})^2,\ \bar{x} = \frac{1}{n}\sum_{h=1}^{n} x_h$$

$$s_x^2(n') = \frac{1}{n'-1}\sum_{h=1}^{n'}(x_h - \bar{x}')^2,\ \bar{x}' = \frac{1}{n'}\sum_{h=1}^{n'} x_h$$

The two-phase sampling version of the exponential estimator for S_y^2 proposed by R. Singh et al. (2011) is

$$t_2 = s_y^2(n)\exp\left[\frac{s_x^2(n') - s_x^2(n)}{s_x^2(n') + s_x^2(n)}\right] \tag{2}$$

Additional auxiliary variables which are highly correlated to the study variable y can be used to enhance the precision of the estimator. Motivated by

Chand (1975), consider a chain ratio-type estimator using information on two auxiliary variables x and z for estimating S_y^2 as

$$t_3 = s_y^2(n)\frac{s_x^2(n')}{s_x^2(n)}\frac{S_z^2}{s_z^2(n')} \tag{3}$$

A modified chain ratio-type estimator for S_y^2 suggested by H. P. Singh, Mathur, and Chandra (2009) is

$$t_4 = s_y^2(n)\frac{s_x^2(n')}{s_x^2(n)}\left\{\frac{S_z^2+\beta_2(z)}{s_z^2(n')+\beta_2(z)}\right\} \tag{4}$$

where

$$s_z^2(n') = \frac{1}{n'-1}\sum_{h=1}^{n'}(z_h-\bar{z}')^2,\ \bar{z}' = \frac{1}{n'}\sum_{h=1}^{n'}z_h$$
$$S_z^2 = \frac{1}{N-1}\sum_{h=1}^{N}(z_h-\bar{Z})^2,\ \bar{Z} = \frac{1}{N}\sum_{h=1}^{N}z_h$$

and $\beta_2(z)$ is the known population coefficient of kurtosis of the variable z. There may be several auxiliary information, which if efficiently utilized can improve the precision of the estimates.

Motivated by the above, consider an unbiased estimator for the population variance S_y^2 of the study variable y using p (non-negative integer constant) auxiliary variables x_i ($i = 1, 2,\ldots, p$) as

$$\mathrm{T_{RK}}(p) = K_1 s_y^2(n) + K_2 s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n')}{s_{x_i}^2(n)} + K_3 s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n)}{s_{x_i}^2(n')} \tag{5}$$

where

$$s_{x_i}^2(n) = \frac{1}{n-1}\sum_{h=1}^{n}(x_{i_h}-\bar{x}_i)^2,\ \bar{x}_i = \frac{1}{n}\sum_{h=1}^{n}x_{i_h},\ i = 1,2,\ldots,p$$

$$s_{x_i}^2(n') = \frac{1}{n'-1}\sum_{h=1}^{n'}\left(x_{i_h} - \bar{x}_i'\right)^2,\ \bar{x}_i' = \frac{1}{n'}\sum_{h=1}^{n'} x_{i_h},\ i = 1, 2, \ldots, p$$

and the K_i (i = 1, 2, 3) are real scalars suitably chosen such that

$$K_1 + pK_2 + pK_3 = 1 \tag{6}$$

Remark 1: The estimator $T_{RK}(p)$ is proposed under the following conditions:

i. The sum $(K_1 + pK_2 + pK_3)$ is one.
ii. The weights of the linear form are chose such that the approximate bias is zero.
iii. The approximate variance of $T_{RK}(p)$ attains minimum.

Properties of the Estimator $T_{RK}(p)$

Noted from equation (5), the proposed estimator $T_{RK}(p)$ is biased for S_y^2. Following Remark 1, it may be made unbiased for up to the first order of approximations. The variance V(.) up to the first order of approximations are derived under large sample approximations using the following transformations:

$$\left.\begin{aligned} &s_y^2(n) = S_y^2(1+e_0),\\ &s_{x_i}^2(n) = S_{x_i}^2(1+e_{1i})\\ &s_{x_i}^2(n') = S_{x_i}^2(1+e_{2i}) \end{aligned}\right\} \text{for } i = 1, 2, \ldots, p$$
$$\mathrm{E}(e_0) = \mathrm{E}(e_{1i}) = \mathrm{E}(e_{2i}) = 0 \quad \forall i = 1, \ldots, p$$

Under the above transformations, the estimator $T_{RK}(p)$ takes the following form:

$$\begin{aligned} T_{RK}(p) = {} & K_1 S_y^2(1+e_0) + K_2 S_y^2(1+e_0)\sum_{i=1}^{p}(1+e_{2i})(1+e_{1i})^{-1} \\ & + K_3 S_y^2(1+e_0)\sum_{i=1}^{p}(1+e_{1i})(1+e_{2i})^{-1} \end{aligned} \tag{7}$$

Hence, the bias and mean square error of the estimator $T_{RK}(p)$ must be derived separately for Cases I and II of the two-phase sampling structure.

Case I

The second phase sample S is drawn as a subsample of the first phase sample S'. In this case, the expected values of the sample statistics are

$$\left.\begin{aligned} &\mathrm{E}\left(e_0^2\right)=f_1C_0^2,\ \mathrm{E}\left(e_{1i}^2\right)=f_1C_i^2,\ \mathrm{E}\left(e_{2i}^2\right)=f_2C_i^2 \\ &\mathrm{E}\left(e_0e_{1i}\right)=f_1\rho_{0i}C_0C_i,\ \mathrm{E}\left(e_0e_{2i}\right)=f_2\rho_{0i}C_0C_i,\ \mathrm{E}\left(e_{1i}e_{2i}\right)=f_2C_i^2 \\ &\mathrm{E}\left(e_{1i}e_{1j}\right)=f_1\rho_{ij}C_iC_j,\ \mathrm{E}\left(e_{2i}e_{2j}\right)=f_2\rho_{ij}C_iC_j,\ \mathrm{E}\left(e_{1i}e_{2j}\right)=f_2\rho_{ij}C_iC_j \end{aligned}\right\} \quad (8)$$

where, for integers $s, t \geq 0$,

$$\mu(i)_{st}=\frac{1}{N}\sum_{h=1}^{N}\left\{\left(y_h-\bar{Y}\right)^s\left(x_{i_h}-\bar{X}_i\right)^t\right\},\ \lambda(i)_{st}=\frac{\mu(i)_{st}}{\sqrt{\mu(i)_{20}^s\,\mu(i)_{02}^t}},$$

$$C_0=\sqrt{\left(\lambda(i)_{40}-1\right)},\ C_i=\sqrt{\left(\lambda(i)_{04}-1\right)}\ ,\ \rho_{0i}=\frac{\left(\lambda(i)_{22}-1\right)}{\sqrt{\left(\lambda(i)_{40}-1\right)\left(\lambda(i)_{04}-1\right)}}$$

$$\mu(ij)_{st}=\frac{1}{N}\sum_{h=1}^{N}\left\{\left(x_{i_h}-\bar{X}_i\right)^s\left(x_{j_h}-\bar{X}_j\right)^t\right\},$$

$$\lambda(ij)_{st}=\frac{\mu(ij)_{st}}{\sqrt{\mu(ij)_{20}^s\,\mu(ij)_{02}^t}},\ \rho_{ij}=\frac{\left(\lambda(ij)_{22}-1\right)}{\sqrt{\left(\lambda(ij)_{40}-1\right)\left(\lambda(ij)_{04}-1\right)}},$$

$$A_{0i}=\rho_{0i}\frac{C_0}{C_i},\ A_{ij}=\rho_{ij}\frac{C_i}{C_j},\ \forall i,j=1,2,\ldots,p,$$

$$\text{and } f_1=\frac{1}{n}-\frac{1}{N},\ f_2=\frac{1}{n'}-\frac{1}{N},\ f_3=\frac{1}{n}-\frac{1}{n'}$$

Expanding the right-hand side of equation (7) in terms of the *e* and using the results from equation (8), the expression of bias and mean square error of the estimator $T_{RK}(p)$ using large sample approximations is

$$\begin{aligned} \mathrm{B}\left[\mathrm{T}_{\mathrm{RK}}(p)\right] &= \mathrm{E}\left[\mathrm{T}_{\mathrm{RK}}(p) - S_y^2\right] \\ &= K_2 f_3 S_y^2 \sum_{i=1}^{p} C_i^2 - \left(K_2 - K_3\right) f_3 S_y^2 \sum_{i=1}^{p} A_{0i} C_i^2 \end{aligned} \tag{9}$$

$$\begin{aligned} \mathrm{M}\left[\mathrm{T}_{\mathrm{RK}}(p)\right] &= \mathrm{E}\left[\mathrm{T}_{\mathrm{RK}}(p) - S_y^2\right]^2 \\ &= S_y^4 \left[f_1 C_0^2 + \alpha^2 f_3 \left\{ \sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_{ij} C_j^2 \right\} - 2\alpha f_3 \sum_{i=1}^{p} A_{0i} C_i^2 \right] \end{aligned} \tag{10}$$

where

$$\alpha = K_2 - K_3 \tag{11}$$

Minimization of the mean square error in equation (10) with respect to α yields its optimum value as

$$\alpha_{\mathrm{opt}} = \frac{\left(\sum_{i=1}^{p} A_{0i} C_i^2\right)^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_{0i} C_i^2} \tag{12}$$

Substituting the optimum value of α in equation (10) we obtain the minimum mean square error of $\mathrm{T}_{\mathrm{RK}}(p)$ as

$$\mathrm{Min.M}\left[\mathrm{T}_{\mathrm{RK}}(p)\right] = S_y^4 \left[f_1 C_0^2 - f_3 \frac{\left(\sum_{i=1}^{p} A_{0i} C_i^2\right)^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_{0i} C_i^2} \right] \tag{13}$$

Further, from equations (11) and (12),

$$\alpha_{\mathrm{opt}} = \left(K_2\right)_{\mathrm{opt}} - \left(K_3\right)_{\mathrm{opt}} = \frac{\sum_{i=1}^{p} A_{0i} C_i^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_{0i} C_i^2} = R(\mathrm{say}) \tag{14}$$

From equations (6) and (14), note that only two equations in three unknowns are not sufficient to find the unique values of the K_i (i = 1, 2, 3). In order to get unique values of the K_i, impose a linear restriction as

$$\mathrm{B}\left[\mathrm{T}_{\mathrm{RK}}\left(p\right)\right]=0 \tag{15}$$

Thus from equation (9),

$$K_2\sum_{i=1}^{p}\left(1-A_{0i}\right)C_i^2+K_3\sum_{i=1}^{p}A_{0i}C_i^2=0 \tag{16}$$

Equations (6), (14), and (16) can be written in matrix form as

$$\begin{pmatrix}1 & p & p\\ 0 & 1 & -1\\ 0 & \sum_{i=1}^{p}\left(1-A_{0i}\right)C_i^2 & \sum_{i=1}^{p}A_{0i}C_i^2\end{pmatrix}\times\begin{pmatrix}K_1\\K_2\\K_3\end{pmatrix}=\begin{pmatrix}1\\R\\0\end{pmatrix} \tag{17}$$

Solving (17), we get the unique values of the K_i as

$$\begin{aligned}\left(K_1\right)_{\mathrm{opt}}&=1-p\left[\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}\left(2A_{0i}-1\right)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\right]\\ \left(K_2\right)_{\mathrm{opt}}&=\frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\\ \left(K_3\right)_{\mathrm{opt}}&=\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}\left(A_{0i}-1\right)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\end{aligned} \tag{18}$$

From equation (18), substituting the values of $(K_1)_{\mathrm{opt}}$, $(K_2)_{\mathrm{opt}}$, and $(K_3)_{\mathrm{opt}}$ in equation (5) yields the optimum unbiased estimator for S_y^2 as

$$\mathrm{T}_{\mathrm{RK}}(p)=\left[1-p\left\{\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(2A_{0i}-1)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\right\}\right]s_y^2(n)$$
$$+\frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n')}{s_{x_i}^2(n)} \qquad (19)$$
$$+\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(A_{0i}-1)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n)}{s_{x_i}^2(n')}$$

whose optimum variance up to the first degree of approximations is given by

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}(p)\right]_{\mathrm{opt}}=S_y^4\left[f_1C_0^2-f_3\frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2}\right] \qquad (20)$$

Case II

When the second-phase sample S is drawn independently of the first-phase sample S'. In this case, the following expected values of the sample statistics are

$$\left.\begin{aligned}&\mathrm{E}\left(e_0^2\right)=f_1C_0^2,\ \mathrm{E}\left(e_{1i}^2\right)=f_1C_i^2,\ \mathrm{E}\left(e_{2i}^2\right)=f_2C_i^2\\&\mathrm{E}\left(e_0e_{1i}\right)=f_1\rho_{0i}C_0C_i,\ \mathrm{E}\left(e_{1i}e_{1j}\right)=f_1\rho_{ij}C_iC_j,\ \mathrm{E}\left(e_{2i}e_{2j}\right)=f_2\rho_{ij}C_iC_j,\\&\mathrm{E}\left(e_0e_{2i}\right)=\mathrm{E}\left(e_{1i}e_{2i}\right)=\mathrm{E}\left(e_{1i}e_{2j}\right)=0\end{aligned}\right\} \qquad (21)$$

Proceeding as in Case I, the optimum unbiased estimator for S_y^2 is obtained as

$$\mathrm{T}_{\mathrm{RK}}(p)=\left[1-\frac{pf_1}{(f_1+f_2)^2}\left\{\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(2f_1A_{0i}-f_3)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\right\}\right]s_y^2(n)$$

$$+\frac{f_1}{(f_1+f_2)^2}\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(f_2+f_1A_{0i})C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n')}{s_{x_i}^2(n)} \qquad (22)$$

$$+\left(\frac{f_1}{f_1+f_2}\right)^2\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(A_{0i}-1)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n)}{s_{x_i}^2(n')}$$

with optimum variance up-to first order of approximations as

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}(p)\right]_{\mathrm{opt}}=S_y^4\left[f_1C_0^2-\frac{f_1^2}{f_1+f_2}\frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2}\right] \qquad (23)$$

Remark 2: It is to be noted from equation (18) that the unique value of the scalars K_i ($i = 1, 2, 3$) involved in estimator depend on unknown population parameters C_0, C_i, ρ_{0i}, and ρ_{ij} ($i, j = 1, 2,\ldots, p$). Thus, to make the estimator practicable, one has to use the guessed or estimated values of these unknown population parameters. Guessed values of population parameters can be obtained either from past data or experience gathered over time; see Murthy (1967), Reddy (1978), and Tracy, Singh, and Singh (1996). If the guessed values are not known then it is advisable to use their respective sample estimates as suggested by Upadhyaya and Singh (1999), H. P. Singh, Chandra, Joarder, and Singh (2007), and Gupta and Shabbir (2008). The minimum variance of the proposed class of estimators remains the same up to the first order of approximations, even if population parameters are replaced by their respective sample estimates.

Empirical Investigations

As p, the number of auxiliary variables, is a non-negative integer, therefore it is not practically possible to deal with the suggested estimator $T_{RK}(p)$ in its general form to carry out the numerical illustrations. Thus, for empirical investigations, consider $T_{RK}(p)$ with $p = 1$ and 2, where the suggested estimator $T_{RK}(p)$ is superior to t_1 and t_2 for $T_{RK}(1)$ (i.e. $p = 1$) and dominates t_3 and t_4 for $p = 2$. The performance of $T_{RK}(1)$ is examined under two different cases of double sampling. The MSEs of the estimators t_1, t_2, t_3, and t_4 and the variance of $T_{RK}(p)$ (for $p = 1, 2$) up to first order of approximations under both the Cases I and II of two-phase sampling set up are presented below.

Case I

$$\mathrm{M}(t_1) = S_y^4\left[f_1 C_0^2 + f_3 C_1^2 \left(1 - 2A_{01}\right)\right]$$

$$\mathrm{M}(t_2) = S_y^4\left[f_1 C_0^2 + \frac{1}{4} f_3 C_1^2 \left(1 - 4A_{01}\right)\right]$$

$$\mathrm{M}(t_3) = S_y^4\left[f_1 C_0^2 + f_3 C_1^2 \left(1 - 2A_{01}\right) + f_2 C_2^2 \left(1 - 2A_{02}\right)\right]$$

$$\mathrm{M}(t_4) = S_y^4\left[f_1 C_0^2 + f_3 C_1^2 \left(1 - 2A_{01}\right) + \theta f_2 C_2^2 \left(\theta - 2A_{02}\right)\right]$$

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}(1)\right] = S_y^4\left[f_1 C_0^2 - f_3 A_{01}^2 C_1^2 \right]$$

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}(2)\right] = S_y^4\left[f_1 C_0^2 - f_3 \frac{\left(A_{01} C_1^2 + A_{02} C_2^2\right)^2}{C_1^2 + C_2^2 + A_{12} C_2^2 + A_{21} C_1^2}\right]$$

where

$$\theta = \frac{S_z^2}{S_z^2 + \beta_2(z)}$$

Case II

$$\mathrm{M}(t_1) = S_y^4\left[f_1C_0^2 + (f_1+f_2)C_1^2\left(1-2\frac{f_1}{f_1+f_2}A_{01}\right)\right]$$

$$\mathrm{M}(t_2) = S_y^4\left[f_1C_0^2 + \frac{1}{4}(f_1+f_2)C_1^2\left(1-4\frac{f_1}{f_1+f_2}A_{01}\right)\right]$$

$$\mathrm{M}(t_3) = S_y^4\left[f_1C_0^2 + (f_1+f_2)C_1^2\left(1-2\frac{f_1}{f_1+f_2}A_{01}\right) + f_2C_2^2\left(1-2A_{12}\right)\right]$$

$$\mathrm{M}(t_4) = S_y^4\left[f_1C_0^2 + (f_1+f_2)C_1^2\left(1-2\frac{f_1}{f_1+f_2}A_{01}\right) + \theta f_2C_2^2\left(\theta-2A_{12}\right)\right]$$

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}(1)\right] = S_y^4\left[f_1C_0^2 - \frac{f_1^2}{f_1+f_2}A_{01}^2C_1^2\right]$$

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}(2)\right] = S_y^4\left[f_1C_0^2 - \frac{f_1^2}{f_1+f_2}\frac{\left(A_{01}C_1^2 + A_{02}C_2^2\right)^2}{C_1^2 + C_2^2 + A_{12}C_2^2 + A_{21}C_1^2}\right]$$

with θ as described above.

Numerical Illustration using Known Natural Populations

Six natural datasets were chosen to elucidate the efficacious performance of the proposed estimator $\mathrm{T}_{\mathrm{RK}}(p)$ (for $p = 1, 2$) over the estimators stated above. The source of the variables y, x, and z and the values of the various parameters are given below.

Population I: Source: Murthy (1967, p. 288).

y: Output.

x: Fixed capital.

z: Number of workers.

Population II: Source: Cochran (1977, p. 182).

y: Food cost.
x: Size of the family.
z: Income.

Population III: Source: Anderson (1958).

y: Head length of second son.
x: Head length of first son.
z: Head breadth of first son.

Population IV: Source: Wang and Chen (2012, p. 39).

y: Volume.
x: Diameter.
z: Height.

Population V: Source: Dobson (1990, p. 192).

y: Survival time.
x: White blood cell count.
z: White blood cell count at page number 74.

Population VI: Source: Sukhatme and Sukhatme (1970, p. 185).

y: Area (acres) under wheat in 1937.
x: Area (acres) under wheat in 1936.
z: Total cultivated area (acres) in 1931.

Table 1. Parametric values of different populations

Population	N	θ	C_0	C_1	C_2	ρ_{01}	ρ_{02}	ρ_{12}
I	80	0.999996	1.1255	1.6065	1.3662	0.7319	0.7940	0.9716
II	33	0.981200	1.0104	1.1780	1.0691	0.1341	0.4630	0.3905
III	25	0.953485	1.3512	1.4295	1.2853	0.5057	0.5683	0.4213
IV	31	0.943500	1.2634	1.2018	1.1962	0.7448	0.0547	0.3256
V	17	0.152800	0.8351	1.4049	1.0818	-0.0144	0.4468	0.5790
VI	34	1.000000	1.5959	1.5105	1.3200	0.6251	0.8007	0.5342

The values of various parameters obtained from above populations are presented in Table 1.

To obtain a tangible idea about the performance of the proposed estimator $T_{RK}(p)$ (for $p = 1, 2$), the percent relative efficiencies (PREs) of $T_{RK}(p)$ (for $p = 1, 2$) and other estimators were computed with respect to the sample variance $s_y^2(n)$, the natural estimator for S_y^2, for both the cases of two-phase sampling set up. The results are demonstrated in Tables 2 and 3.

The PRE of an estimator $T_{RK}(p)$ with respect to sample variance estimator s_y^2 is defined as

$$\text{PRE} = \frac{\text{V}\left(s_y^2\right)}{\text{V}\left[\text{T}_{\text{RK}}\left(p\right)\right]_{\text{opt}}} \times 100 \tag{24}$$

Numerical Example using Artificially Generated Population

Three sets of independent random numbers were generated of size N ($N = 100$), x_k', y_k', and z_k' ($k = 1, 2, 3, \ldots, N$) from a standard normal distribution via R. Motivated by the artificial data set generation techniques adopted by S. Singh and Deo (2003) and S. Singh, Joarder, and Tracy (2001), the following transformed variables of U were generated with the values of $\sigma_y^2 = 100$, $\mu_y = 40$, $\sigma_x^2 = 225$, $\mu_x = 50$, $\sigma_z^2 = 25$, and $\mu_z = 30$ as

$$y_k = \mu_y + \sigma_y\left[\rho_{xy}x_k' + \left(\sqrt{1-\rho_{xy}^2}\right)y_k'\right],\ x_k = \mu_x + \sigma_y x_k',$$
$$\text{and } z_k = \mu_z + \sigma_z\left[\rho_{xz}x_k' + \left(\sqrt{1-\rho_{xz}^2}\right)z_k'\right]$$

PREs of different estimators for fixed and varying values of ρ_{xy} and ρ_{xz} are presented in Tables 3 and 4, respectively.

Table 2. PREs of different estimators

Population			Percent Relative Efficiency											
Pop. I			Case I						Case II					
N	*n'*	*n*	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
80	65	45	103.796	160.387	160.447	120.674	120.675	170.217	*	162.396	170.389	100.937	100.937	182.593
		40	104.167	170.012	170.085	116.933	116.933	182.212	*	171.764	177.066	101.913	101.913	191.046
		30	104.691	185.605	185.703	112.068	112.068	202.155	*	186.854	188.867	103.313	103.313	206.274
	50	35	102.853	139.961	139.996	131.523	131.523	145.523	*	142.380	157.539	*	*	166.643
		25	103.931	163.758	163.823	119.287	119.287	174.391	*	165.682	172.679	101.290	101.290	185.479
		20	104.341	174.910	174.991	115.265	115.265	188.407	*	176.515	180.635	102.376	102.376	195.613
Pop. II			Case I						Case II					
N	*n'*	*n*	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
33	25	12	*	*	101.492	*	*	111.007	*	*	101.545	*	*	111.432
		10	*	*	101.574	*	*	111.665	*	*	101.605	*	*	111.923
		8	*	*	101.642	*	*	112.224	*	*	101.66	*	*	112.369
	15	8	*	*	101.121	*	*	108.079	*	*	101.317	*	*	109.611
		6	*	*	101.337	*	*	109.768	*	*	101.441	*	*	110.595
		4	*	*	101.525	*	*	111.267	*	*	101.568	*	*	111.622
Pop. III			Case I						Case II					
N	*n'*	*n*	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
25	20	12	*	124.425	124.489	100.282	101.028	144.897	*	123.551	126.228	*	*	148.651
		10	*	127.01	127.083	*	*	150.529	*	126.351	128.074	*	*	152.734
		7	*	129.934	130.017	*	*	157.146	*	129.531	130.39	*	*	158.008
	15	8	*	121.231	121.286	102.2	103.257	138.205	*	120.107	124.172	*	*	144.22
		6	*	125.23	125.297	*	100.499	146.629	*	124.422	126.78	*	*	149.87
		4	*	128.665	128.743	*	*	154.24	*	128.149	129.352	*	*	155.623

Note: "*" indicates no gain, i.e., PRE is less than 100

Table 2, continued.

Population			Percent Relative Efficiency											
Pop. IV			Case I						Case II					
N	n'	n	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
31	17	12	132.59	130.114	136.282	*	*	113.144	104.343	157.393	157.471	*	*	118.941
		10	145.21	141.475	150.88	100.041	103.811	117.253	118.68	164.905	166.253	103.718	106.548	121.049
		8	157.601	152.472	165.527	116.059	119.744	120.88	133.876	171.675	175.826	119.654	122.393	123.179
	12	8	129.891	127.662	133.2	*	*	112.203	110.422	160.73	161.166	*	*	119.847
		6	146.535	142.658	152.432	101.654	105.424	117.659	120.25	165.656	167.229	105.331	108.158	121.274
		5	155.339	150.476	162.83	112.972	116.688	120.245	131.013	170.48	174.001	116.597	119.363	122.786
Pop. V			Case I						Case II					
N	n'	n	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
17	12	8	*	*	100.013	*	*	102.832	*	*	104.438	*	*	100.098
		7	*	*	100.015	*	*	103.197	*	*	104.721	*	*	100.104
		6	*	*	100.016	*	*	103.498	*	*	104.981	*	*	100.11
	10	7	*	*	100.011	*	*	102.281	*	*	104.067	*	*	100.09
		6	*	*	100.013	*	*	102.779	*	*	104.399	*	*	100.097
		5	*	*	100.015	*	*	103.197	*	*	104.721	*	*	100.104
Pop. VI			Case I						Case II					
N	n'	n	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
34	25	12	130.044	141.943	145.778	155.47	155.47	209.612	103.4	143.687	143.736	108.245	108.245	202.714
		10	132.338	145.463	149.733	151.613	151.613	223.762	112.559	147.867	148.505	116.48	116.48	219.25
		8	134.343	148.581	153.251	148.495	148.495	237.318	120.683	151.21	152.888	123.628	123.628	235.875
	15	7	123.927	132.792	135.581	167.614	167.614	177.625	*	138.968	139.116	*	*	188.047
		6	126.494	136.592	139.801	162.149	162.149	190.144	104.639	144.281	144.371	109.37	109.37	204.829
		4	131.394	144.008	148.096	153.161	153.161	217.773	115.766	149.225	150.218	119.319	119.319	225.573

Note: "*" indicates no gain, i.e., PRE is less than 100

Table 3. PREs of different estimators under artificially generated populations for ρ_{xy} = 0.7 and ρ_{xz} = 0.5

Artificial Population			Estimators						
			Case I						
N	*n*'	*n*	$s_y^2(n)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
100	80	55	100	*	108.8652	109.8181	*	*	108.0435
		45	100	*	110.2873	111.4091	*	*	109.3225
		40	100	*	110.8303	112.0177	*	*	109.8100
	70	50	100	*	107.1820	107.9408	*	*	106.5256
		40	100	*	109.1417	110.1271	*	*	108.2923
		30	100	*	110.5859	111.7437	*	*	109.5906
			Case II						
N	*n*'	*n*	$s_y^2(n)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
100	80	55	100	*	105.3662	110.9395	*	*	108.9443
		45	100	*	107.8582	111.9666	*	*	109.7676
		40	100	*	108.8234	112.4033	*	*	110.1168
	70	50	100	*	102.4825	109.9041	*	*	108.1116
		40	100	*	105.8466	111.1271	*	*	109.0948
		30	100	*	108.3879	112.2033	*	*	109.9569

Note: "*" indicates no gain, i.e., PRE is less than 100

Table 4. PREs of Different estimators for varying values of ρ_{xy} and ρ_{xz}

		Case I Estimators					
ρ_{xy}	ρ_{xz}	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
0.8	0.8	101.983	116.671	116.696	*	101.536	117.440
	0.6	126.096	121.277	127.007	109.818	115.747	118.407
	0.4	115.223	117.736	119.551	*	*	109.180
	0.2	*	119.547	119.551	*	*	111.733
0.5	0.8	*	*	100.349	*	*	100.390
	0.6	*	102.123	103.171	*	*	101.590
	0.4	*	*	100.159	*	*	100.227
	0.2	*	*	102.017	*	*	100.300
0.2	0.8	*	*	100.188	*	*	100.573
	0.6	*	*	100.033	*	*	100.025
	0.4	*	*	100.035	*	*	100.351
	0.2	*	*	100.289	*	*	101.920

Note: "*" indicates no gain, i.e., PRE is less than 100

Table 4, continued.

		Case II Estimators					
ρ_{xy}	ρ_{xz}	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
0.8	0.8	*	119.103	132.247	*	*	133.885
	0.6	106.535	156.666	156.841	*	101.806	136.045
	0.4	*	136.901	138.644	*	*	116.728
	0.2	*	118.359	138.644	*	*	121.799
0.5	0.8	*	*	100.596	*	*	100.666
	0.6	*	*	105.528	*	*	102.740
	0.4	*	*	100.272	*	*	100.387
	0.2	*	*	103.488	*	*	100.511
0.2	0.8	*	*	100.321	*	*	100.981
	0.6	*	*	100.057	*	*	100.044
	0.4	*	*	100.059	*	*	100.600
	0.2	*	*	100.493	*	*	103.318

Note: "*" indicates no gain, i.e., PRE is less than 100

Conclusion

For natural population datasets, Table 2 exhibits that, under different structures of two-phase sampling set up, our suggested estimator $T_{RK}(p)$ (for $p = 1$ and 2) is superior to the existing one under its respective optimality condition and also preferable in general situations. For fixed n' (first-phase sample size), the PRE of the proposed estimator is increasing with decreasing values of n (second-phase sample size), i.e. the smaller the second phase sample, the more efficiency in $T_{RK}(p)$ will be achieved, which reduces the cost of the survey.

For the artificially generated data set, the results compiled in Table 3 indicate the proposed methodology yielded impressive gains in efficiency over the existing methods, and same behavior in efficiency of $T_{RK}(p)$ was reflected, indicating the proposed methodology is cost-effective.

It can also be observed from Table 4 that if several populations are generated artificially for various combinations of values of ρ_{xy} and ρ_{xz}, our proposed methodology is always preferable over the existing one. The proposition of the estimator in the present study is justified as it unifies several desirable results including unbiased and efficient estimation strategy, and may be recommended for practical applications.

References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York, NY: John Wiley & Sons.

Chand, L. (1975). *Some ratio type estimators based on two or more auxiliary variables* (Unpublished doctoral dissertation). Iowa State University, Ames, IA. Retrieved from: http://lib.dr.iastate.edu/rtd/5190/

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: Wiley.

Das, A. K., & Tripathi, T. P. (1978). Use of auxiliary information in estimating the finite population variance. *Sankhyā, Series C, 40*(2), 139-148.

Dobson, A. J. (1990). *An introduction to generalized linear models* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9781420057683

Gupta, S., & Shabbir, J. (2008). On improvement in estimating the population mean in simple random sampling. *Journal of Applied Statistics, 35*(5), 559-566. doi: 10.1080/02664760701835839

Isaki, C. T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association, 78*(381), 117-123. doi: 10.2307/2287117

Murthy, M. N. (1967). *Sampling theory and methods*. Calcutta, India: Statistical Publishing Society.

Prasad, B., & Singh, H. P. (1990). Some improved ratio type estimators of finite population variance in sample surveys. *Communications in Statistics – Theory and Methods, 19*(3), 1127-1139. doi: 10.1080/03610929008830251

Prasad, B., & Singh, H. P. (1992). Unbiased estimators of finite population variance using auxiliary information in sample surveys. *Communications in Statistics – Theory and Methods, 21*(5), 1367-1376. doi: 10.1080/03610929208830852

Reddy, V. N. (1978). A study on the use of prior knowledge on certain population parameters in estimation. *Sankhyā, Series C, 40*(1), 29-37.

Singh, H. P., Chandra, P., Joarder, A. H., & Singh, S. (2007). Family of estimators of mean, ratio and product of a finite population using random nonresponse. *TEST, 16*(3), 565-597. doi: 10.1007/s11749-006-0020-z

Singh, H. P., Mathur, N., & Chandra, P. (2009). A chain type estimator for population variance using two auxiliary variables in two-phase sampling. *Statistics in Transition New Series, 10*(1), 75-84.

Singh, R., Chauhan, P., Sawan, N., & Smarandache, F. (2011). Improved exponential estimator for population variance using two auxiliary variables. *Italian Journal of Pure and Applied Mathematics, 28-2011*, 101-108.

Singh, R. K. (1983). Estimation of finite population variance using ratio and product method of estimation. *Biometrical Journal, 25*(2), 193-200.

Singh, S., & Deo, B. (2003). Imputation by power transformation. *Statistical Papers, 44*(4), 555-579. doi: 10.1007/bf02926010

Singh, S., & Joarder, A. H. (1998). Estimation of finite population variance using random non-response in survey sampling. *Metrika, 47*(1), 241-249. doi: 10.1007/bf02742876

Singh, S., Joarder, A. H., & Tracy, D. S. (2001). Median estimation using double sampling. *Australian & New Zealand Journal of Statistics, 43*(1), 33-46. doi: 10.1111/1467-842x.00153

Srivastava, S. K., & Jhaji, H. S. (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhyā, Series C, 42*(12), 87-96.

Sukhatme, P. V., & Sukhatme, B. V. (1970). *Sampling theory of surveys with application*. Ames, IA: Iowa State University Press.

Tailor, R., & Sharma, B. (2012). Modified estimators of population variance in presence of auxiliary information. *Statistics in Transition New Series, 13*(1), 37-46.

Tracy, D. S., Singh, H. P., & Singh, R. (1996). An alternative to the ratio-cum-product estimator in sample surveys. *Journal of Statistical Planning and Inference, 53*(3), 375-387. doi: 10.1016/0378-3758(95)00136-0

Tripathi, T. P., Singh, H. P., & Upadhyaya, L. N. (1988). A generalized method of estimation in double sampling. *Journal of the Indian Statistical Association, 26*, 91-101.

Upadhyaya, L. N., & Singh, H. P. (1983). Use of auxiliary information in the estimation of population variance. *Mathematical Forum, 6*(2), 33-36.

Upadhyaya, L. N., & Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal, 41*(5), 627-636. doi: 10.1002/(SICI)1521-4036(199909)41:5<627::AID-BIMJ627>3.0.CO;2-W

Wang, Y. W., & Chen, H.-J. (2012). Use of percentiles and Z-scores in anthropometry. In V. R. Preedy (Ed.), *Hand book of anthropometry: Physical*

measures of human form in health and disease (pp. 29-48). New York, NY: Springer. doi: 10.1007/978-1-4419-1788-1_2

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 179-194.
doi: 10.22237/jmasm/1493597400

ISSN 1538 − 9472

Analysis of Robust Parameter Designs

Tak K. Mak
Concordia University
Montreal, Québec

Fassil Nebebe
Concordia University
Montreal, Québec

The analysis of robust parameter design is discussed via a model incorporating mean-variance relationship which, when ignored as in the classical regression approach, can be problematic. The model is also capable of alleviating the difficulties of the regression approach in the search of the minimum variance occurring region.

Keywords: Graphical log-linear models, contingency tables, decomposable models, hierarchical log-linear models

Introduction

As part of their efforts in quality improvement, manufacturers strive to design products that are capable of functioning optimally under a wide range of environmental conditions. Instead of using more expensive parts or components, a more cost-effective means is to look for settings of design factors that would achieve this quality robustness. Specifically, this involves finding design settings that would minimize variance while being on target. In this regard, robust parameter designs have been widely used in the industry to determine the optimal settings of these design factors (Khuri & Mukhopadhyay, 2010; Robinson, Borror, & Myers, 2004). In robust parameter designs, design of experiment techniques are used to obtain data that are subsequently analyzed to explore the relationship between the quality characteristics and the levels of the design factors (Choi & Allen, 2009).

Taguchi advocated the use of crossed array designs and suggested a convenient analysis using signal to noise ratio. However, the limitations of an analysis based on the "signal to noise" ratio have been pointed out by many researchers (Box, 1988; Barreau, Chassagnon, Kobi, & Seibilia, 1999). Various methods of analysis have been proposed in robust parameter designs. Vining and Myers (1990) suggested a dual response surface methodology in which the

T. K. Mak is a professor in the department of Supply Chain & B. T. M. Email him at: tak.mak@concordia.ca.

(primary) response surface of standard deviation is minimized subject to a target constraint based on a (secondary) response surface of the mean (see also Chan & Mak, 1995).

Another common approach exploits the possible existence of mean-variance relationship to achieve simpler variance minimization computationally. With this method, the variance is assumed to be a product of a function of the mean and a "Performance Measure Independent of Adjustment" (PerMIA) (Box, 1988, p. 2). The PerMIA is a function of a proper subset of design factors (control factors) and the complement of this subset constitutes the subset of adjustment factors which influence variance only through its presence in the mean. Because of this variance factorization it is possible to minimize variance through the unconstrained minimization of the PerMIA, which is then followed by the searching of the levels of the adjustment variables to attain the desired target value.

Both the dual response surface methodology and the PerMIA approach conduct an analysis based on the sample variance or standard deviation calculated from replicates at each treatment combination. For the crossed array design, the sample variances are calculated from observations in the outer arrays that are crossed with the treatment combinations or inner arrays in the experiment. The sample variances calculated from the outer arrays do not constitute estimates of variances obtained from random samples. In crossed array or combined array designs, the noise factors which occur randomly during the lifetime usage of the product are controlled and have known values in the experiment (Welch, Yu, Kang, & Sacks, 1990; Shoemaker, Tsui, & Wu, 1991; Mak & Nebebe, 2005).

Thus in the design stage, roles of the design and noise factors are indistinguishable. In the analysis of such designs, a regression function is first fitted, from which the variance function is derived with respect to the randomness of the noise factors. Variance minimization can then be conducted based on the inherent variance function from this regression modeling approach. This regression analysis is conceptually simple and exploits the quality characteristic and noise factor relationships to achieve possibly greater efficiency. However, this regression approach has two issues to be properly addressed: First, it does not accommodate the possible dependency of the variance on the mean. Mak and Nebebe (2004) demonstrated with an example that settings determined by the regression approach can yield a variance that is substantially higher than the actual attainable optimal variance. They also proposed a new model that incorporates the mean-variance relation and includes the regression model as a special case. Second, the form of the variance function is not flexible and completely determined by the formulated regression model based on the

interactions between the control and noise variables (O'Donnell & Vining, 1997). It does not permit the formulation of a simple linear relationship between the variance and the design factors and the variance model has to be at least of the second order. Unfortunately, as seen in the simulation studies in this paper, this second order variance model in the regression approach is usually inadequate. The aforementioned issues in the regression approach are addressed in this study, and some practical recommendations are made in the light of the simulation result.

Methodology

Modeling Mean-Variance Relationship

Denote by y the quality characteristic of interest. Let $X_1,\ldots, X_p$ be p design factors influencing y. Suppose that there are q noise factors $Z_1,\ldots, Z_q$, the levels of which are controlled in the experimenting stage. Mak and Nebebe (2004) proposed the following model for analyzing robust parameter designs:

$$y = \mu_y\left(\mathbf{x},\boldsymbol{\beta}\right) + \sqrt{\mathrm{V}_\lambda\left(\mu_y\left(\mathbf{x},\boldsymbol{\beta}\right)\right)}\left(\mathrm{h}\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right)+e\right) \tag{1}$$

where $\mathbf{x} = (X_1,\ldots, X_p)'$, $\mathbf{z} = (Z_1,\ldots, Z_q)'$, e is the error term with variance σ_e^2, $\mu_y(\mathbf{x}, \boldsymbol{\beta})$ is the conditional mean of y given $\mathbf{x}$ (with respect to the distribution of $\mathbf{z}$ and e), and $\boldsymbol{\beta}$ is the vector of regression parameters. Furthermore, $\mathbf{x}_{(1)} \subseteq \mathbf{x}$ is a subset of "control factors", and $\mathrm{V}_\lambda(\mu_y)$ is a scalar function with parameter λ specifying the dependence of the variance on the mean. It is required that $\mathrm{V}_\lambda(\mu_y) \equiv 1$ for a certain λ, say $\lambda = 0$. Because $\mathrm{E}(y) = \mu_y(\mathbf{x}, \boldsymbol{\beta})$, it follows that $\mathrm{E}(\mathrm{h}(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta})) = 0$, where the expectation $\mathrm{E}(\cdot)$ is taken with respect to the distribution of $\mathbf{z}$. It follows from (1) that

$$\mathrm{Var}\left(y\right) = \mathrm{V}_\lambda\left(\mu_y\right)\left\{\mathrm{Var}\left(\mathrm{h}\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right)\right)+\sigma_e^2\right\}$$

where Var is the variance operator taken with respect to the distribution of $\mathbf{z}$ and e, μ_y is written in place of $\mu_y(\mathbf{x}, \boldsymbol{\beta})$ when there is no possibility of ambiguity. Because $\mathrm{V}_\lambda(\mu_y) \equiv 1$ when $\lambda = 0$, Mak and Nebebe's model includes the situation where there is no mean-variance relation as a special case. It is clear that the PerMIA is given by $\mathrm{Var}(\mathrm{h}(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}))+\sigma_e^2$, and $\mathbf{x}_{(2)} = \mathbf{x} \setminus \mathbf{x}_{(1)}$ is the vector of adjustment factors that influence the variance through its presence in the mean. Thus to

minimize variance around a target, one can choose levels of $\mathbf{x}_{(1)}$ to minimize the PerMIA and then adjust the levels of $\mathbf{x}_{(2)}$ to attain the desired target mean. With the use of PerMIA, only unconstrained minimization is needed and a change in the target value requires only readjustment of the levels of the adjustment factors (Box, 1988). In the next section, we give an algorithm for computing iteratively estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ given λ.

Box (1988) proposed a transformation approach for designs with replicates which can be easily extended to crossed array or combined array designs. Specifically, it is assumed that there exists a transformation T_λ such that $T_\lambda(y) = m(\mathbf{x}, \boldsymbol{\beta}) + h(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta}) + e$, eliminating the dependency of the variance on the mean on the transformed scale. Thus $y = T_\lambda^{-1}\left\{m\left(\mathbf{x},\boldsymbol{\beta}\right)+h\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right)+e\right\}$ and to terms of the linear order, we have approximately,

$$y = T_\lambda^{-1}\left(m\left(\mathbf{x},\boldsymbol{\beta}\right)\right)+\left.\frac{d\left(T_\lambda^{-1}\left(u\right)\right)}{du}\right|_{u=m\left(\mathbf{x},\boldsymbol{\beta}\right)}\left\{h\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right)+e\right\}$$

which is in the form of (1). Thus we have seen that Mak and Nebebe's model is approximately equivalent to Box's (1988) transformation approach. The analysis conducted on the transformed scale using Box's approach has to be followed by an "aim-off" analysis in order that variance be minimized on the original metric.

Determining λ and the identification of $\mathbf{x}_{(1)}$

From (1), if

$$y_* = \frac{y-\mu_y}{\sqrt{V_\lambda\left(\mu_y\right)}} = h\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right)+e \tag{2}$$

then there is an ordinary regression model with homogeneous errors. However, if the λ used on the left is different from the true value of λ, say λ_0, then

$$y_* = \frac{y-\mu_y}{\sqrt{V_\lambda\left(\mu_y\right)}} = \frac{\sqrt{V_{\lambda_0}\left(\mu_y\right)}}{\sqrt{V_\lambda\left(\mu_y\right)}}h\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right)+\frac{\sqrt{V_{\lambda_0}\left(\mu_y\right)}}{\sqrt{V_\lambda\left(\mu_y\right)}}e$$

Table 1. Simulated data and mean, variance calculations

				z_1	-1	1	1	-1				
				z_2	-1	1	-1	1	True		Sample	
x_1	x_2	x_3	x_4	z_3	-1	-1	1	1	mean	var	mean	var
-1	0	0	0		22.2	25.1	22.5	28.0	26.2	2.5	24.5	5.5
-1	1	1	1		26.0	28.4	25.6	26.9	29.8	4.3	26.7	1.1
-1	-1	-1	-1		19.1	20.1	17.2	24.9	20.8	1.0	20.3	8.1
0	-1	0	1		9.0	10.5	10.5	10.1	9.8	0.7	10.0	0.4
0	0	1	-1		24.5	36.9	40.0	22.9	30.2	62.7	31.1	56.0
0	1	-1	0		20.4	31.0	28.7	21.1	26.0	34.7	25.3	21.6
1	-1	1	0		25.8	26.9	24.7	28.7	28.2	3.4	26.6	2.2
1	0	-1	1		19.5	18.9	19.5	20.3	19.5	0.8	19.5	0.2
1	1	0	-1		23.6	25.0	22.6	28.5	26.1	2.5	24.9	4.9

Because $\mu_y = E(y) = \mu_y(\mathbf{x}, \boldsymbol{\beta})$ is a function of possibly all the design factors, the adjustment factors would also appear to have some influences on the variable y_*. Thus the relationship between y_* and the adjustment variables will be zero only when the true λ is used in (2) so that the explained variation by each of these adjustment variables should attain its minimum at a value around the true λ. This fact could be exploited to determine the value of λ approximately. To illustrate this, consider a set of simulated data from a crossed array design involving four design factors and three noise factors. The inner and outer arrays are, respectively, L9 and L4 arrays so that the experiment consists of (9)(4) = 36 experimental runs crossed between the inner and outer arrays. For each experimental run, the quality characteristic y is simulated using the model

$$y = \mu_y + \mu_y^2\left(\left(0.0024 + 0.003u_2\left(X_1\right)\right)Z_1 + 0.001Z_2 - 0.002Z_3 + e\right)$$

where $u_2(X)$ is the quadratic orthogonal polynomial $2 - 3X$, $u_1(X) = X$ is the linear orthogonal polynomial, and $\sigma_e = 0.003$. The levels of the design and noise factors in the design and the simulated data are given in Table 1, along with the true mean μ_y of y for each inner array used to simulate the data.

The mean of y for each of the nine inner arrays is simply estimated by the mean of the y values from the corresponding crossed outer array. $V_\lambda\left(\mu_\lambda\right) = \mu_y^\lambda$ and the true λ is equal to 4. Note the h function can be approximated by a quadratic function in $\mathbf{z}$ but the second order terms vanish since $E(h(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta}) = 0$. If we also retain only up to the quadratic effects of the X_i on the variance, then the h function is approximately a linear combination

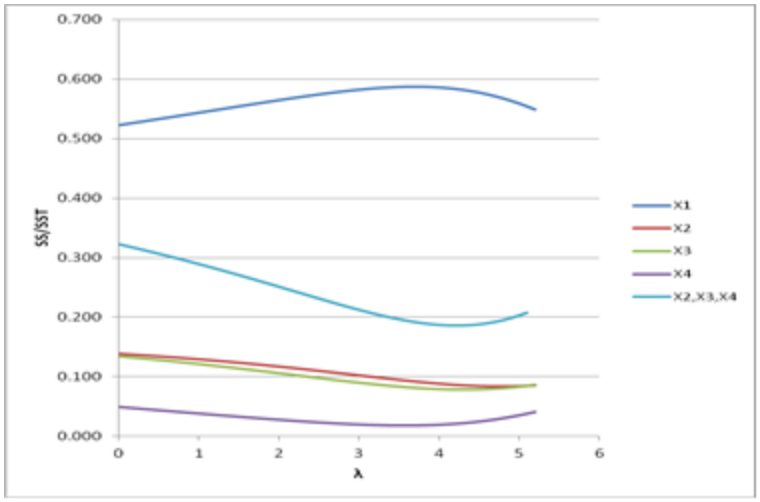


Figure 1. The proportion of sum of squares plot

$$h(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta}) = \sum_{j=1}^{3} \theta_j Z_j + \sum_{i=1}^{4} \sum_{j=1}^{3} \left[\theta_{ij1} u_1(X_i) Z_j + \theta_{ij2} u_2(X_i) Z_j \right]$$

of the terms. Z_j, $u_1(X_i)Z_j$, $u_2(X_i)Z_j$, $i = 1,\ldots, 4$, $j = 1, 2, 3$. Thus, even if the functional form of the h function is not exactly known, the suggested method can still be used to determine λ, though the value is only approximately unbiased. Also, as with any estimation procedure, the suggested method will not yield the true value of λ due to the randomness of the error term e. In the present example, the linear combination of effects described above is used as the h function in fitting the regression model (2) in order to determine λ. For any given λ, the linear regression model (2) can therefore also be fitted to the data yielding the total sum of squares SS(X_i) corresponding to each design factor X_i (i.e. total of the sum of squares for $u_1(X_i)Z_j$, $u_2(X_i)Z_j$, $j = 1, 2, 3$), for $i = 1,\ldots, 4$. A graphical plot is presented in Figure 1 of P(X_i) = SS(X_i)/SST against λ for $i = 1,\ldots, 4$.

It is clear that X_1 is the only control variable affecting variance and X_2, X_3, and X_4 are adjustment variables. Furthermore, the value of P(X_i) is smallest when λ is equal to 4.7, 4.3, and 3.6 for X_2, X_3, and X_4, respectively. The proportion of sum of squares

$$\mathrm{P} = \sum_{i=1,2,3,4} \frac{\mathrm{SS}(X_i)}{\mathrm{SST}}$$

corresponding to all the adjustment variables X_2, X_3, and X_4 is also plotted against λ in Figure 1. The value of P attains a minimum at $\lambda = 4.2$, which indicates that, collectively, the observed relationship between y_* and the variables X_2, X_3, and X_4 are lowest when λ is close to 4. Thus in the present example, the suggested method determines quite accurately the true value of λ.

Estimation of Parameters

Consider the iterative estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ for given λ. Suppose there are n experimental runs in the experiment. Let $X_{1i},\ldots, X_{pi}$ and $Z_{1i},\ldots, Z_{qi}$ be the levels of the design and noise factors, respectively, in the i^{th} run. Let y_i be the observed quality characteristic and $\mu_{yi} = E(y_i) = \mu_y(\mathbf{x}_i, \boldsymbol{\beta})$, where now $\mathbf{x}_i = (X_{1i},\ldots, X_{pi})'$. The computing of the estimates calls an external algorithm denoted, say, by WLS($\mathbf{y}$, $\mathbf{X}$, r(·), $\mathbf{w}$, $\mathbf{p}$). The input arguments $\mathbf{y}$, $\mathbf{X}$, r, and $\mathbf{w}$ of WLS represent, respectively, the array of values of the dependent variable, the design matrix, the regression function, and the array of weights. The array $\mathbf{p}$ holds the output weighted least squares estimates of the regression parameters. The algorithm for computing the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is given below.

Step 0. Initialize and save the starting values of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ in the arrays $\mathbf{b}_0$ and $\mathbf{f}_0$.

Step 1. For $i = 1$ to n
- i. Read the values of $\mathbf{x}_i$, $\mathbf{x}_{(1)i}$, $\mathbf{z}_i$ into the i^{th} row of two-dimensional arrays **XA**, **XA1**, **ZA**
- ii. Next
- iii. Read the values of y into a one-dimensional array $\mathbf{y}$

Step 2. For $i = 1$ to n
- i. Let $m = \mu(\mathbf{x}_i, \mathbf{b}_0)$. Here $\mathbf{x}_i$ is from the i^{th} row of **XA**
- ii. Let $ys(i) = \left[y(i) - m \right] \Big/ \sqrt{V_\lambda(m)}$
- iii. Let $w(i) = 1$
- iv. Read $\mathbf{x}_{(1)i}$, $\mathbf{z}_i$ into the i^{th} row of a two-dimensional array **XZA**
- v. Next

Step 3. Call WLS($\mathbf{ys}$, **XZA**, h(·), $\mathbf{w}$, $\mathbf{f}_1$). Here the regression function h is $h(\mathbf{x}_{(1)i}, \mathbf{z}_i, \boldsymbol{\theta})$

Step 4. For $i = 1$ to n
- i. Let $m = \mu(\mathbf{x}_i, \mathbf{b}_0)$. Here $\mathbf{x}_i$ is from the i^{th} row of **XA**

ii. Let $u(i) = y(i) - \sqrt{V_\lambda(m)}\, h\left(\mathbf{x}_{(1)}, \mathbf{z}_i, \mathbf{f}_1\right)$. Here $\mathbf{x}_{(1)}$, $\mathbf{z}_i$ are from the i^{th} row of **XA1**, **ZA**

iii. Let $w(i) = 1/V_\lambda(m)$

iv. Next

Step 5. Call WLS(**u**, **XA**, μ(·), **w**, $\mathbf{b}_1$). Here the regression function μ is μ($\mathbf{x}_i$, **β**)

Step 6. If $\mathbf{b}_1$ and $\mathbf{f}_1$ differ from respectively $\mathbf{b}_0$ and $\mathbf{f}_0$ by less than certain prescribed small values

i. Then

- Stop the program

ii. Else

- Let $\mathbf{b}_0 = \mathbf{b}_1$
- Let $\mathbf{f}_0 = \mathbf{f}_1$
- Go to Step 2

iii. End if

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ be the final values of $\mathbf{b}_1$ and $\mathbf{f}_1$ obtained from the iterative procedure. The estimate $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\right)$ is a solution of the system of equations in **β** and **θ**:

$$\sum \left(\frac{y_i - \mu_y\left(\mathbf{x}_i, \boldsymbol{\beta}\right)}{\sqrt{V_\lambda\left(\mu_y\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right)}} - h\left(\mathbf{x}_{(1)i}, \mathbf{z}_i, \boldsymbol{\theta}\right) \right) \frac{\partial h\left(\mathbf{x}_{(1)i}, \mathbf{z}_i, \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = 0,$$

$$\sum \frac{1}{V_\lambda\left(\mu_y\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right)} \left(y_i - \sqrt{V_\lambda\left(\mu_y\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right)}\, h\left(\mathbf{x}_{(1)i}, \mathbf{z}_i, \boldsymbol{\theta}\right) - \mu_y\left(\mathbf{x}_i, \boldsymbol{\beta}\right) \right) \frac{\partial \mu_y\left(\mathbf{x}_i, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = 0$$

Because the left sides of this system of equations have zero expected values, the estimator $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\right)$ is consistent for (**β**, **θ**).

Another method of estimation is the familiar maximum likelihood approach in which the error term is assumed to be normally distributed. The maximum likelihood estimators do not have a closed form and must be obtained numerically. One possibility is to use a generic search optimization algorithm, such as the simplex search, that does not require first or second order derivatives. It is however quite well-known in the regression literature that the estimation of the mean could be more adversely affected when the variance function is mis-specified. In this regard, the dual response surface methodology in which separate

regression models are fitted with the sample means and standard deviations calculated from the replicates (or observations from the outer arrays) may be more robust in the estimation of the mean to the misspecification of the variance component. However, for experiments with few experimental runs and a large number of factors, a quadratic response surface for the variance may not be feasible. The estimation procedure proposed in this paper may also be adversely affected by the mis-specification of the h function as it also appears in the second estimation equation for estimating the **β**.

Results: Extensions of the Classical Regression Approach

To simplify the notation and the discussion of the classical regression (CR) approach and its possible ramification, assume the variance is functionally not dependent on the mean. In the presence of a variance-mean relationship, the analyses proposed here can be readily generalized to incorporate such a relationship using the method above. In the classical regression approach, the mean function $\mu_y(\mathbf{x}, \boldsymbol{\beta})$ is usually assumed to be quadratic in form (without cross-product terms between factors in some designs). The dependence of the variance on the control factors is typically introduced into the model by incorporating in the function $h(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta})$ cross-product terms between the terms in the quadratic mean model and the noise factors. For example, for the orthogonal inner array L9 used in Vandenbrande's (1998) experiment, the mean function involves the saturated model:

$$\mu_y\left(\mathbf{x},\boldsymbol{\beta}\right)=\mu+\beta_{11}u_1\left(X_1\right)+\beta_{12}u_2\left(X_1\right)+\ldots+\beta_{41}u_1\left(X_4\right)+\beta_{42}u_2\left(X_4\right)$$

and

$$\begin{aligned}\mathrm{h}\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right)&=\sum_{j=1}^{3}\theta_j Z_j+\sum_{i=1}^{4}\sum_{j=1}^{3}\left[\theta_{ij1}u_1\left(X_i\right)z_j+\theta_{ij2}u_2\left(X_i\right)Z_j\right]\\&=\sum_{j=1}^{3}\left[\theta_j+\sum_{i=1}^{4}\theta_{ij1}u_1\left(X_i\right)+\theta_{ij2}u_2\left(X_i\right)\right]Z_j\end{aligned}$$

Consequently,

$$\mathrm{Var}(y) = \sigma_e^2 + \sum_{j=1}^{3}\left[\theta_j + \sum_{i=1}^{4}\theta_{ij1}u_1(X_i) + \theta_{ij2}u_2(X_i)\right]^2 \sigma_{z_j}^2$$

where, without loss of generality, the variances of Z_1, Z_2, and Z_3 can be taken to be unity. Thus under the classical regression model the fourth order of a control factor may be involved. If this interaction model is approximately valid and the variance function can be reasonably approximated by a quadratic function, then $\theta_{ij2} = 0$ for all i and j and the following reduced model can be considered:

$$\mathrm{h}\left(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta}\right) = \sum_{j=1}^{3}\theta_j Z_j + \sum_{i=1}^{4}\sum_{j=1}^{3}\theta_{ij1}u_1(X_i)Z_j$$

The variance function for this reduced model becomes:

$$\mathrm{Var}(y) = \sigma_e^2 + \sum_{j=1}^{3}\left(\theta_j + \sum_{i=1}^{4}\theta_{ij1}X_i\right)^2$$

because $u_1(x) = x$ and $\sigma_{z_j}^2 = 1$. It is clear that a major drawback of the interaction model is that the variance can never assume the form of a linear function of the control factors and is therefore not appropriate for designs used in the steepest descent stage for locating the region containing the optimal variance solutions.

Proposed here is a generalization of the classical model to attain greater efficiency and flexibility in handling a wider range of applications. Consider the following functional form of the function h:

$$\mathrm{h}\left(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta}\right) = \sum_{j}\mathrm{sign}\left(\psi_j\right)|\psi_j|^{\gamma} Z_j \tag{3}$$

indexed by γ, where ψ_j is a function of $\mathbf{x}_{(1)}$ with a vector parameter $\boldsymbol{\theta}_{(j)}$. In most applications, we could choose ψ_j to be a quadratic function of $\mathbf{x}_{(1)}$. When $\gamma = 1$ and ψ_j is a quadratic function, the model clearly becomes the classical regression model. The case $\gamma = 0.5$ is also of special interest since it yields a linear or quadratic approximation to the variance function depending on whether the functional forms chosen for ψ_j are linear or quadratic.

For the analysis of crossed array designs, the dual response surface method (Vining & Myers, 1990) is a serious competitor to the classical regression

approach. In the dual response surface approach, a quadratic response surface is additionally postulated with the sample variance (or standard deviation) as the response variable. It is argued that the calculation of the sample variance for each inner array using the observations from the crossed outer array is not entirely appropriate, because they do not constitute a random sample. If

$$y = \mu_y(\mathbf{x}, \boldsymbol{\beta}) + \mathrm{h}\left(\mathbf{x}_{(1)}, \mathbf{z}, \boldsymbol{\theta}\right) + e$$

for any functional form of h, it is not difficult to show that an unbiased estimator of the variance for each inner array is indeed given by

$$a^{-1}\sum(y - \bar{y})^2 - a^{-1}\sigma_\varepsilon^2$$

where a is the dimension of the outer arrays (number of observations in the outer array). Note that the divisor in the first term is a not $a - 1$. Thus if σ_ε is small, an approximately unbiased estimator of the variance for each inner array is given by the simple estimator $v_i = a^{-1}\sum(y - \bar{y})^2$.

Simulation

Simulation studies were conducted to compare the performances of the following four variance estimators discussed above:

1. *Response surface method (RSM)*. The value v_i is used as the response value of the i^{th} inner array. A quadratic model is then fitted to this variance response surface.
2. *The classical regression approach (CR)*.
3. *The classical regression approach leading to a quadratic variance function (CRQV)*. This model includes only the cross-product terms between the linear effects of the design factors and the noise factors in the function h, which gives rise to a quadratic variance function as explained in the previous section.
4. *The generalized regression approach (GR)* with $\gamma = 0.5$ in (3).

Again, the design studied in Mak and Nebebe (2004) and Vandenbrande (1998) were used. The means and variances used to simulate the data are given in Table 2. The true model used to generate the *y* observations is:

Table 2. True means and variance of the model used to simulate the data

x_1	x_2	x_3	x_4	True mean	True variance
-1	0	0	0	41.15	18
-1	1	1	1	44.80	18
-1	-1	-1	-1	35.83	18
0	-1	0	1	24.83	7
0	0	1	-1	45.18	7
0	1	-1	0	41.03	7
1	-1	1	0	43.20	8
1	0	-1	1	34.47	8
1	1	0	-1	41.05	8

$$y = \mu_y\left(\mathbf{x},\boldsymbol{\beta}\right) + \mathrm{h}\left(\mathbf{x}_{(1)},\mathbf{z},\boldsymbol{\theta}\right) + e$$

where h is given by (3) with $\gamma = 0.5$. Here, for the simplicity of comparisons, the factor x_1 is the only control factor appearing in the functions ψ_j:

$$\psi_1 = -5 + 3X_1 + 2u_2\left(X_1\right),\ \psi_2 = 1 - X_1,\ \psi_3 = -1 + X_1$$

The standard deviation of the normally distributed error term is 2. Two hundred samples were simulated from the true model in the Monte Carlo studies. For each simulated sample, the four methods RSM, CR, CRQV, and GR are each used to fit a variance function. Table 3 summarizes the results for estimating the variances of *y* for X_1 = -1.5, -1.0, -0.5, 0, 0.5, 1.0, 1.5. In reporting the simulation results, we calculate both an estimate's relative bias (RB, defined as (mean of variance estimate – true variance)/true variance) and the coefficient of variation (CV, defined as SD of variance estimate/Mean of variance estimate).

It is seen that in most cases, the CR and the CRQV approaches can be heavily biased (with RB greater than 15%) even for X_1 within the boundary of the experimental region. The bias is particularly severe if the two approaches are used for extrapolating variances (X_1 = -1.5 and 1.5). The two approaches have in general about the same CV in estimating variances. These CV of variance estimates are also comparable to those of the GR approaches (with the exception

of the case $X_1 = 1.5$ where the GR approach has a considerably smaller CV) which in general has smaller biases. The RSM approach has about the same biases as GR and far smaller biases than CRQV in most cases. These observations suggest that the classical approach can be very inadequate even if the true model is reasonably approximated by a quadratic function.

Table 3. Expected values and standard deviations of variance estimates obtained by simulations

		RSM		GR		CRQV		CR	
x_1	True variance	Mean RB	SD CV	Mean RB	SD CV	Mean RB	SD CV	Mean RB	SD CV
-1.5	28	26.08	8.74	26.88	6.25	19.25	4.90	41.47	13.02
		-0.069	0.335	-0.040	0.233	-0.313	0.254	0.481	0.314
-1.0	18	16.69	3.94	16.86	3.49	15.00	3.23	17.51	3.64
		-0.073	0.236	-0.064	0.207	-0.167	0.215	-0.027	0.208
-0.5	11	10.25	2.80	10.15	2.32	11.80	2.51	8.55	2.16
		-0.068	0.273	-0.077	0.228	0.073	0.213	-0.223	0.253
0.0	7	5.96	2.10	5.99	1.59	8.77	1.96	5.33	1.45
		-0.148	0.351	-0.145	0.265	0.253	0.224	-0.239	0.271
0.5	6	4.84	1.93	5.02	1.61	6.88	1.81	4.63	1.36
		-0.194	0.398	-0.164	0.321	0.146	0.263	-0.229	0.294
1.0	8	7.03	3.28	7.19	3.03	5.97	2.29	7.09	2.90
		-0.122	0.466	-0.101	0.421	-0.253	0.383	-0.114	0.409
1.5	15	12.14	6.91	13.57	5.27	5.86	3.05	16.63	8.61
		-0.190	0.569	-0.096	0.388	-0.609	0.521	0.109	0.518

The simulation studies shed some light on the performance of the different methods in practice. Of the four approaches, the RSM is the only one that does not rely on the knowledge of the functional form of the function h. In fact, it does not even model variance involving the noise variables **Z** controlled in the experiment. It simply approximates the variance function directly with a linear or quadratic function of the design variables. Consequently, it does not suffer from the same potential model misspecification (of the variance) experienced by the other methods. This is consistent with the simulation results as the bias of RSM is seen to be generally smaller than those based on the regression approach. However, since the fitting relies on the sample standard deviations based on repeated observations, which in general have greater sampling variability, this robustness is achieved at the expense of an inflated variance of variance estimates – of the four approaches, it has substantially higher CV. For the regression approaches (GR, CR, CRQV), the variance parameters in $\boldsymbol{\theta}$ are more efficiently

estimated as the regression coefficients of a mean regression model and therefore, as observed in the simulation studies, have smaller variances than RSM. Thus in practice, the regression approaches may be preferred, but caution must be taken to ensure the validity of the model, especially the functional form of the function h. The quadratic variance function of the CRQV approach is actually in the form of the square of a linear function and therefore does not have the same effectiveness in approximating h as a general quadratic function. In this regard, the extension suggested in the previous section, provides a more flexible and effective means of approximating the true h function, as demonstrated in the simulation studies where GR has considerably smaller biases than CR and CRQV in most cases.

Conclusion

Mak and Nebebe (2004) demonstrated the importance of the incorporation of the mean-variance relationship, if it exists, in analyzing crossed or combined array designs. They also proposed a model generalizing the traditional method of analysis. In this paper, we proposed a simple method of determining an appropriate mean-variance relation to be used in the model. An estimation procedure is also proposed for the model. With a numerical example, the advantages of Mak and Nebebe's model is demonstrated in terms of variance minimization. In terms of robustness of mean estimation to mis-specification of the variance function, the dual response surface methodology is also appealing, though it has other limitations. It might also be interesting to modify the proposed estimation by modifying the second estimating equation so that the estimation of the regression parameter is still consistent but less adversely affected by model misspecification.

The model proposed by Mak and Nebebe (2004) assumed an error term with homogeneous variances. In analyzing combined array designs, Engel and Huele (1996) considered a model in which the error terms have heterogeneous error variances which are functions of some of the design factors (they however assume $V_\lambda(\mu_y) \equiv 1$). This generalization may also be incorporated in Mak and Nebebe's model and the iterative estimation method suggested will then have to be modified accordingly, using traditional methods of regression analysis with heterogeneous variances. However, when the noise factors have already accounted for the majority of the unconditional variance of the quality characteristic so that the error term is in general small, this modification may not yield substantial practical differences.

Acknowledgements

The authors are grateful to the referee for the helpful comments. This research was supported by NSERC-GRF grant (N01374) and VPRGS seed funding (VS0141) of Concordia University.

References

Barreau, A., Chassagnon, R., Kobi, A., & Seibilia, B. (1999). Taguchi's parameter design: An improved alternative approach. *53rd Annual Quality Congress Proceedings* (pp. 400-404). Milwaukee, WI: American Society for Quality.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformation. *Technometrics, 30*(1), 1-31. doi: 10.2307/1270311

Chan, L. K., & Mak, T. K. (1995). A regression approach for discovering small variation around a target. *Applied Statistics, 44*(3), 369-377. doi: 10.2307/2986043

Choi, H. J., & Allen, J. K. (2009). A metamodeling approach for uncertainty analysis of nondeterministic systems. *Journal of Mechanical Design, 131*(4), 041008. doi: 10.1115/1.3087565

Engel, J., & Huele, A. F. (1996). A generalized linear modeling approach to robust design. *Technometrics, 39*(4), 365-373. doi: 10.2307/1271307

Khuri, A. I., & Mukhopadhyay, S. (2010). Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(2), 128-149. doi: 10.1002/wics.73

Mak, T. K., & Nebebe, F. (2004). Modeling problems in parameter designs. *Journal of Statistical Research, 38*, 155-166.

Mak, T. K., & Nebebe, F. (2005). Estimation of process variances in robust parameter designs. *Journal of Modern Applied Statistical Methods, 4*(2), 394-401. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol4/iss2/5/

O'Donnell, E. M., & Vining, G. G. (1997). Mean squared error of prediction approach to the analysis of a combined array. *Journal of Applied Statistics, 24*(6), 733-746. doi: 10.1080/02664769723468

Robinson, T. J., Borror, C. M., & Myers, M. H. (2004). Robust parameter design: A review. *Quality and Reliability Engineering International, 20*(1), 81-101. doi: 10.1002/qre.602

Shoemaker, A. C., Tsui, K. L., & Wu, C. F. (1991). Economical experimentation methods for robust design. *Technometrics, 33*(4), 415-427. doi: 10.2307/1269414

Vandenbrande, W. (1998). SPC in paint application: Mission impossible. *52nd Annual Quality Congress Proceedings* (pp. 708-715). Milwaukee, WI: American Society for Quality.

Vining, G. G., & Myers, R. H. (1990). Combining Taguchi and response surface philosophies – A dual response approach. *Journal of Quality Technology, 22*(1), 38-45.

Welch, W. J., Yu, T. K., Kang, S. M., & Sacks, J. (1990). Computer experiments for quality control by parameter design. *Journal of Quality Technology, 22*(1), 15-22.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 195-232.
doi: 10.22237/jmasm/1493597460

ISSN 1538 − 9472

Robustness and Power Comparison of the Mood-Westenberg and Siegel-Tukey Tests

Linda C. Lowenstein
Washington University in St. Louis
St. Louis, MO

Shlomo S. Sawilowsky
Wayne State University
Detroit, MI

The Mood-Westenberg and Siegel-Tukey tests were examined to determine their robustness with respect to Type-I error for detecting variance changes when their assumptions of equal means were slightly violated, a condition that approaches the Behrens-Fisher problem. Monte Carlo methods were used via 34,606 variations of sample sizes, α levels, distributions/data sets, treatments modeled as a change in scale, and treatments modeled as a shift in means. The Siegel-Tukey was the more robust, and was able to handle a more diverse set of conditions.

Keywords: Behrens-Fisher, Mood-Westenberg, Siegel-Tukey

Introduction

"Heteroscedasticity, refers to situations where two or more of the variances are unequal" (Wilcox, 1996, p. 174). The applied statistical literature is vast on how poorly the *t* and *F* tests perform under this condition. For instance, it has been demonstrated that small sample sizes, unequal sample sizes, and one-tailed tests can be problematic for the *t*-test with respect to heteroscedasticity and non-normal data (Sawilowsky & Blair, 1992; Wilcox, 1996; Sawilowsky, 2002). With respect to the Analysis of Variance (ANOVA) *F* test, the problem is even worse (Brown & Forsythe, 1974; Rogan & Keselman, 1977; Tomarken & Serlin, 1986). Wilcox (1996) stated "our hope is that any problem associated with unequal variances might diminish when there are more than two groups, but the reverse seems to be true" (p. 180). Referring to the ratio (*R*) of standard deviation between groups in a survey of educational studies, Wilcox (1996) "found that that estimates of *R* are

Linda C. Lowenstein is a Postdoctoral Researcher in Applied Mathematics and the corresponding author. Email her at: lclowenstein@gmail.com. Dr. Sawilowsky is a Professor in the College of Education and the founding editor of this journal. Email him at professorshlomo@gmail.com.

often higher than 4" (p. 180; see Wilcox, 1989), noting *R*'s as large as 11 were observed in real world data applications.

Keppel and Wickens (2004) noted "the actual significance level could appreciably exceed the nominal α level when the group variances were unequal. Under these circumstances, we need a way to adjust or modify our analysis" (p. 152). Hence, inflated Type-I errors lead to pronouncements of the statistical significance of nonsense treatments.

Under the truth of the null hypothesis, the counter-argument is having equal means with unequal variance is unrealistic (see, e.g., Sawilowsky, 2002). "That is, this situation will never arise in practice because if the variances are unequal, surely the means are unequal, in which case a Type-I error is not an issue" (Wilcox, 1996, p. 180). The condition of unequal variances between groups is known as the Behrens-Fisher problem, named after the work of W. V. Behrens (1929) and Sir Ronald A. Fisher (1935, 1939) who developed the first expression and approximate solution. Sawilowsky (2002) noted the Behrens-Fisher problem "arises in testing the difference between two means with a *t* test when the ratio of variances of the two populations from which the data were sampled is not equal to one" (p. 461), and of course expands to layouts with more than two groups.

When the null hypothesis is false, another problem with heteroscedasticity is the *t*, *F*, and other parametric tests' concomitant lack of comparative statistical power. Wilcox (1996) mentioned "there is evidence that problems with Type-I errors with unequal variances reflect undesirable power properties even under normality (Wilcox, Charlin, & Thompson, 1986; Wilcox, 1995)" (p. 180), noting "the power curve might be unusually flat in a region near the null hypothesis (Wilcox, 1995)" especially when the data are skewed (Wilcox, 1996, p. 181). There are situations where the null hypothesis is false, yet the probability of rejecting the null hypothesis is less than α. In this case, small but possibly important treatment effects might be missed.

Sawilowsky and Fahoome (2003) noted non-homogeneity renders most rank-based non-parametric tests even more so ineffective. For example, the Wilcoxon Rank Sum test (Wilcoxon, 1945), which is three to four times more powerful than the t test under common conditions of non-normality due to skew, fares even worse when the treatment impacts scale. Similarly, Sawilowsky (2002) noted "for the case of $K > 2$, Feir-Walsh and Toothaker (1974) and Keselman, Rogan, and Feir-Walsh (1977) found the Kruskal-Wallis test (Kruskal & Wallis, 1952) and expected normal scores test (McSweeney & Penfield, 1969) to be 'substantially affected by inhomogeneity of variance'" (p. 463).

Change in Scale

There are no exact solutions to the Behrens-Fisher problem. According to Wilcox (1996) and Sawilowsky (2002), the non-parametric Yuen solution (Yuen, 1974), with various modifications, is considered as one of the best approximate solutions. Moreover, methods designed for the purpose of detecting scale or variance changes between sample groups with regard to the level of heteroscedasticity necessary to invoke the Behrens-Fisher problem have been generally overlooked in the applied statistical literature. With respect to the often-cited classical Hartley's (1950) *F*-statistic for determining dispersion (variance) differences as a preliminary test, for example, Sawilowsky (2002) noted the deleterious nature of sequential testing that increases the Type-I error rate. Keppel and Wickens (2004) noted the additional problem of non-normality can greatly impact that *F*-statistic for variance difference detection:

> Unfortunately, in spite of its simplicity and of the fact that it is provided by many packaged computer programs, the *F* max statistic is unsatisfactory. Its sampling distribution, as reflected in the Pearson-Hartley tables, is extremely sensitive to the assumption that the scores have a normal distribution. (p. 150)

According to Neave and Worthington (1988), there were no satisfactory nonparametric tests that could determine the potential of unequal variances irrespective of whether there was also a shift in location. They noted the Mood-Westenberg dispersion test (Westenberg, 1948; Mood, 1950), a non-parametric test based on quartile location and Fisher exact probabilities, determined differences in variances under the assumption that the means of two samples are equal, but stopped short of recommending it as a preliminary test for detecting the Behrens-Fisher condition.

Similarly, Neave and Worthington (1988) noted the Siegel-Tukey test (Siegel & Tukey, 1960), another ordinal non-parametric test based on rankings and Mann-Whitney-*U* probabilities, assumes roughly equal means/medians for detecting variance differences between groups. They bemoaned the absence of detection methods for this condition:

> Several attempts have been made to solve the problem, but all resulting tests suffer from being rather un-powerful or not truly distribution-free or both….It is particularly unfortunate that there appears to be no good distribution-free solution to this problem since

> several researchers have shown that non-normality can upset the behavior of the F-statistic to a very considerable extent. (p.135)

The question arises, therefore, if there are no tests that can detect the occurrence of different variances irrespective of means, then how can it be known if heteroscedasticity or the Behrens-Fisher problem arises so as to be alerted to the need to subsequently apply any of the myriad approximate solutions?

Purpose of the Study

There are no early warning or detection systems indicating the Behrens-Fisher condition exists. The Mood-Westenberg and Siegel-Tukey tests appear promising to fill that need in the statistical repertoire in applied data analysis. In the two group layout, both tests assume equal means (or medians) and $\mu_1 = \mu_2$ (or $\theta_1 = \theta_2$). The null hypothesis (H_0) is the variances are equal. The alternative hypothesis (H_A) is that the variances are not equal. The purpose of this study, therefore, is to examine via Monte Carlo methods their Type-I error rates and comparative statistical power properties as the treatment condition approaches the Behrens-Fisher problem, in order to determine if either test can be used as an early warning.

Methodology

Monte Carlo Methods

An Absoft Pro Fortran (version 14.0.4) program with the IMSL Fortran Numerical Library (version 7.0) was coded to randomly select and assign values to simulated control and treatment groups through sampling with replacement. Rangen 2.0 subroutine (Fahoome, 2002), a 90/95 update to the Fortran 77 version (Blair, 1987), was used to generate pseudo-random numbers from the normal and theoretical distributions. Realpops subroutine 2.0 (Sawilowsky, Blair, & Micceri, 1990) was used to generate pseudo-random samples obtained from real education and psychology populations.

For the Mood-Westenberg code, duplicates found in the control (A) and treatment groups (B) were coded to layout the groups as ABABABAB until all duplicates were accounted for; this method was selected as reasonable because this pattern appears to be unbiased for both groups (the pattern could favor either A or B in the extreme quarters depending upon the random variates sampled). Algorithm AS 62 (Dinneen & Blakesley, 1973) was used to calculate the Mann-

Whitney exact probabilities for the Siegel-Tukey test.[1] When sorting was required, the Recursive Fortran 95 quicksort routine that sorts real numbers into ascending numerical order was used.[2]

There were 34,606 combinations of study parameter conditions employed, based on 11 sample sizes, two α levels (0.05, 0.01) (four levels, including 0.025 and 0.005 were calculated and reviewed in preliminary testing), 11 mathematical distributions and real world data sets, 11 variance changes and 13 small means shifts. Independent sample sizes included (n_1, n_2) = (5, 5); (5, 15); (10, 10); (10, 30); (15, 45); (20, 20); (30, 30); (30, 90); (45, 45); (65, 65); (90, 90). They were generated from three theoretical distributions (normal, exponential, uniform), and eight real world education and psychology data sets identified by Micceri (1986, 1989). The data sets were described as smooth symmetric, extreme asymmetric (growth), extreme asymmetric (decline), extreme bimodality, multimodality and lumpy, discrete mass at zero, discrete mass at zero with gap, and digit preference (see Sawilowsky & Blair, 1992). The use of real data sets in addition to data generated from mathematical models was deemed important in rigorous systematic studies by Bradley (1978) and many others.

Next, the means and variances were modified, beginning with no treatment effect via equal means to establish baseline results. Then, treatment effects of location shifts were gradually increased in small magnitudes, thus increasingly violating the statistical assumption of both tests. Type-I (identifying a variance change when none occurred) and Type-II (not finding a true variance change) error rates under the violations were compared to the counterfactual conditions of equal means.

Type-I and -II Errors

In order to determine robustness measures with respect to Type-I and -II errors, the long-run average rejection rates were calculated after executing 100,000 iterations for each study condition. A counter was incremented for statistically significant iterations. The counter totals were reported as rejection percentages (counter total/100,000). Thus, the long-run averages for the p rejection rate, β rejection rate, and power levels $(1 - \beta)$ were determined.

[1] Additional code was provided by Miller, retrieved from http://lib.stat.cmu.edu/apstat/62
[2] Quicksort routine algorithm provided by Rew with additions from Brainard, retrieved from http://www.fortran.com/qsort_c.f95

Robustness Results

A robust test maintains Type-I and -II error rates in light of assumption violations. Bradley's (1978) liberal limits for Type-I errors of $0.5\alpha \leq$ Type-I error $\leq 1.5\alpha$ was adopted.

Asymptotic and exact probabilities were invoked for each test during preliminary testing. For the Mood-Westenberg test, the Chi-squared (asymptotic) and Fisher exact probabilities were selected. For the Siegel-Tukey test, *Z*-scores (asymptotic) and Mann-Whitney (exact) probabilities were selected. Based on the results for the primary testing, only the asymptotic probabilities were reported because the two probabilities for each statistic were found to track closely to each other. Two α levels, 0.05 and 0.01, were reported during the primary testing (four levels, including 0.025 and 0.005, were calculated and reviewed in preliminary testing).

Simulating Location Shifts and Scale Changes

A treatment was modeled as a shift in location, by multiplying a constant c = 0.01-0.12 (0.01) by the distribution's σ. For example, the standard deviation of the smooth symmetric data set was 4.91. Therefore, a treatment effect of $0.1\sigma = 0.491$ was added to the treatment variates. Cohen (1988) suggested $0.2(\sigma)$ represents a small treatment effect, $0.5(\sigma)$ a moderate treatment effect, and $0.8(\sigma)$ a large treatment effect. On the basis of personal communications with Cohen, Sawilowsky (2009) updated Cohen's de facto standards to also define $d(0.01)$ = very small, $d(1.2)$ = very large, and $d(2.0)$ = huge. The focus of this study, based on Sawilowsky's (2009) standard, was to review only small shifts ($c << 0.2$), and therefore the effect sizes of shift in location selected were 0-0.12σ (0.01), $d = 0$ representing the baseline.

A treatment was modeled as a change in scale by multiplying a constant scale shift of $K = 1 – 3.5$ (0.25) by the random variates of the treatment group after they were centered around zero for both groups by subtracting the distribution mean from the variates; this sets the standard deviation of the control group, over the long run, to approach a normal curve having a variance of 1. Heteroscedasticity is simulated when *R*, representing the variance ratio difference between the treatment group and the control group, is not equal to 1. K^2, the new simulated variance of the treatment group, is the ratio difference, *R*, between the post-test treatment and control groups.

It was expected that with ratio variance differences from 1.56 ($K = 1.25$) to 12.25 ($K = 3.5$) (with *K* increments of 0.25 for *K*), the alternative hypothesis (H_1)

would be accepted. When the ratio of the variances between the treatment and control groups was equal to 1 ($K = 1$), the condition of equal variances, then the null hypothesis (H_0) was expected to be retained (i.e., fail to reject). These variance ratio differences are consistent with Brown and Forsythe (1974), who reported standard deviation ratio differences of 3 and found concomitant unacceptably high Type-I error rates, and Wilcox (1989), who surveyed the literature and found estimates of standard deviation ratio differences are often higher than 4, and sometimes even as large as 11.

Results

Simulating No Research Treatment Effects with Equal Means Assumption in Place

Demonstration of Adequacy of Algorithms used in this Simulation: Type-I error for Normal Distribution, Means and Variances are Equal

To demonstrate the adequacy of the algorithms used in this simulation, preliminary testing with data sampled from the Gaussian distribution, with equal mean and variances, was performed for all of sample sizes (Table 1). The minimum and maximum asymptotic upper tail rejection rates for α set at 0.05, 0.025, 0.01, and 0.005 for Mood-Westenberg (Chi-squared) were 0.022-0.080, 0.008-0.033, 0.004-0.033, and 0.000-0.016 respectively. For the Siegel-Tukey (*Z*-scores) they were 0.044-0.058, 0.016-0.027, 0.004-0.010, and 0.000-0.005, respectively. The exact rates tracked close to the associated asymptotic probabilities for both statistics. Exact rates for Mood-Westenberg (Fisher exact) were 0.016-0.072; 0.008-0.033; 0.000-0.020; and 0.000-0.008, and for Siegel-Tukey (Mann-Whitney-*U*) were 0.044-0.050; 0.016-0.025; 0.008-0.010; and 0.004-0.005. The rejection range was larger for Mood-Westenberg. Additional testing for all equal sample sizes (n_1, n_2) = (5, 5) to (200, 200) yielded robust rates for both statistics (Table 2).

For all sample sizes and α levels, Siegel-Tukey's rejection rates for asymptotic and exact probabilities tracked closer to nominal α as compared with the performance of the Mood-Westenberg Chi-squared and Fisher exact probabilities. It appeared that the latter test's Type-I error rates were dependent on the sample size, and it tracked in an unusual and repeating saw-tooth-like pattern as equal sample sizes were increased by 1 from (5, 5) to (200, 200) at 10,000 iterations (Figures 1 and 2).

The Mood-Westenberg Type-I Fisher exact error rates were occasionally nearly as high as 10% when nominal α was 5%, and 2.4% when nominal α was 1%. Thus, the Mood-Westenberg was observed as an inconsistent test because it did not fit the expected pattern for the Type-I rejection rates to approach nominal α level and remain steadfast as the sample size increased. Instead, it moved in and out of threshold defining robustness as the sample sizes increased. This may be due to the instability of the sampling distribution of the median. See Figures 3 and 4 for Siegel-Tukey results.

Type-I Error: All Distributions/Data Sets, Means and Variances are Equal

At large and equal sample sizes ((45, 45) and above), both statistical tests generally demonstrated robust Type-I rates for the distributions and data sets. Conservative non-robust rate exceptions were noted for discrete mass zero with gap, extreme asymmetric decay, and extreme bimodal data sets (Table 3). However, these conservative non-robust rates suggested unlikely pronouncements of false positives when determining variance change in research settings; hence, at this initial stage, each statistic remained viable candidates to provide robust and powerful heteroscedasticity detection with large and equal sample sizes.

With respect to smaller and unequal sample sizes, Mood-Westenberg demonstrated both liberal and conservative non-robust rates for the distributions/data sets while Siegel-Tukey maintained the same robust rates (and conservatively non-robust for the three data sets mentioned above in Table 3) at all sample sizes except for the smallest sample size of (5, 5) where a few more non-robust conservative rates surfaced for other distributions/data sets at α below 5%. At this point, Siegel-Tukey appeared a more consistent statistic for small and unequal sample sizes with respect to Type-I rates.

Type-II Error: All Distributions/Data Sets, Means are Equal and Variances Change (Classical Behrens-Fisher)

For this phase of testing, in order to provide more stability for Mood-Westenberg, the testing occurred only with the large sample size (90, 90) to observe effects of variance changes simulated with the constant K = 1.25-3.5 (0.25). Both statistics were powerful (73-100%) for data sampled from the conservatively non-robust data sets discrete mass zero/gap, extreme asymptotic decay, and extreme bimodal, starting with the smallest variance change when K = 1.25 (Table 4; grey shaded area = 100% power). As to be expected, each statistic demonstrated increases in power as the α levels and variance ratio increased. Strong power for these data sets, with conservative Type-I rates, continued to affirm both statistics as potential

detection tools; these statistics did not lack for power with these data sets. Siegel-Tukey demonstrated consistent power for these data sets at or above 99% while Mood-Westenberg recorded the same and lower power rates for extreme bimodal (73-90%) when $K = 1.25$.

For the other data sets and distribution at sample size (90, 90) (previously all shown to demonstrate robust Type-I error rates), power was lower as compared to the conservatively non-robust data sets mentioned above, yet still good, for both test statistics, particularly for $K = 1.5$ and above. For Mood-Westenberg, power increased dramatically and quickly, doubling or tripling as variance changed from $K = 1.25$-1.5 (Table 5) for these other data sets/distributions. For Siegel-Tukey, the power also increased quickly, but not as dramatically as Mood-Westenberg because the Siegel-Tukey power rates started off higher at lower K constants.

In general, both statistics demonstrated power approaching 40% or higher early on ($K = 1.25$-1.5, larger α). Siegel-Tukey demonstrated power levels equal to or greater than Mood-Westenberg, sometimes 20-40% higher than Mood-Westenberg with smaller variance changes, as demonstrated in Table 4. For instance, at the smallest change of $K = 1.25$, $\alpha = 0.05$, Siegel-Tukey's power rate for smooth symmetric asymptotic was 0.550 compared to Mood-Westenberg at .165. When α equaled 0.01, Siegel-Tukey's rate was 0.288 as compared to Mood-Westenberg's rate at 0.061. When the variance change level was $K = 1.5$ (Table 5), most α levels yielded power of 40-100%, generally, for all distributions and data sets, for both statistics.

The Siegel Tukey asymptotic and exact probabilities (at $\alpha = 0.05$, 0.025, 0.01, and 0.005) consistently demonstrated equal or greater power rates than the Mood-Westenberg probabilities at every comparison point (α and K's) with all distributions/data sets. Both probability measures for Siegel-Tukey quickly approached 100% power, generally arriving with $K = 2$-2.25 (Table 6); Mood-Westenberg arrived at near 100% with $K = 2.75$-3.0. Siegel-Tukey reached power of nearly 90% and above at all α levels at $K = 1.75$, whereas Mood-Westenberg did not reach these levels until $K = 2.25$ (Table 6). As to be expected, power increased for both statistics as variance change and α levels increased, and therefore these preliminary tests demonstrated that each statistic is robust and powerful, in general, when their mutual assumptions of equal means/medians in place. However, Siegel-Tukey generally appeared more powerful than Mood-Westenberg after this testing phase.

Simulating Research Treatment Effects by Violating the Assumption of Equal Means

At this point, attention was turned to the primary focus of the study: would the Mood-Westenberg and the Siegel-Tukey tests remain robust with respect to Type-I and Type-II rejection rates under conditions of simulated treatment effects (i.e., the means began to shift slightly, violating the statistical assumptions). Preliminary testing results of 10,000 means shifts from 0.00001 to 0.1 (0.00001) suggested an appropriate mean shift range, useful for testing, would be 0.01-0.12 (0.01).

To determine the properties for each statistic after sampling from the thousands of combination of populations, sample sizes, means shifts, variance change, and α levels, it would be necessary to review all output, particularly with respect to the smaller and unequal sample sizes. However, general conclusions are made and presented here for both statistics, with respect to whether the mathematical distributions and real-world data sets could be characterized as a normal type distribution (e.g., unimodal shape, asymptotic light tails, symmetric about the means) or not. Normal type distributions are discussed as a group and include: normal, digit preference, discrete mass zero, smooth symmetric, and uniform. Non-normal type distributions, discussed as a group, include: extreme asymmetric growth, extreme asymmetric decay, extreme bimodal, and discrete mass zero with gap. Having demonstrated unique outcomes, exponential and multi-modal lumpy are discussed separately.

With minor exceptions for the exponential and multi-modal lumpy, general conclusions for the distributions and data sets were not greatly affected by the range of the tested means shift levels 0.01-0.12 (0.01); therefore, conclusions for particular distributions and data sets will generally hold for all of the tested means shift levels, especially for larger sample sizes and α levels of 0.05. When robustness was present, larger α levels (0.05), larger and equal sample sizes and larger variance change levels rendered testing measurements more robust and powerful for each distribution and data set.

Type-I Rejection Rates: For All Distributions/Data Sets, Variances are Equal

The statistics were first tested with slight means shifts, $0.01(\sigma)$-$0.12(\sigma)$ (0.01), when simulating post-test equal variance outcomes. Typical results are noted in Table 7 for sample size (90, 90) and mean shift at $c = 0.06$. The expectation was that nominal α rejection rates would hold when the means began to shift. Mood-Westenberg, for most normal type distributions (e.g., digit preference, normal, smooth symmetric, uni), particularly for large sample sizes (i.e., (20, 20);

(30, 30); (30, 90)), maintained generally robust (and conservative non-robust) rejection rates at all of the tested means shifts with some slightly liberal rate exceptions at some small and small/unequal sample sizes or sometimes at 1% α. As noted with sample size (90, 90), in Table 7, the normal type discrete mass zero, sometimes demonstrated small liberal, non-robust rates but robust rejection rates were noted for many other sample sizes, particularly when nominal α was 5%. However, analyzing non-normal distributions (asymmetric growth, discrete mass zero with gap, extreme asymmetric decay, extreme bimodal), Mood-Westenberg, for both asymptotic and exact probabilities at the large sample size (90, 90), calculated many extremely liberally non-robust rejection rates even at the smallest incremental level of 0.01. The test results from data sampled from multi-modal lumpy demonstrated liberal non-robust rejection rates generally at and above means shift $c = 0.09$ for some sample sizes, such as (90, 90), and was robust for many other sample sizes. Results from data sampled from the exponential distribution demonstrated robust rates up to means shifts of 0.06 when, for instance, for sample size (65, 65) or (90, 90) (Table 7), for nominal α below 2.5%, the rejection rates started to trend above nominal α levels in the liberal direction, increasing in slight liberalness with each increase in means shift. Starting with mean shift $c = 0.07$ and above, under Mood-Westenberg, the test results demonstrated that the exponential distribution was liberally non-robust at all α levels for sample size (90, 90). Other sample sizes for exponential also reflected this pattern. Generally, the non-robust Mood-Westenberg results for the exponential distribution were in the liberal direction.

With respect to the Siegel-Tukey statistic, at sample size (90, 90) and mean shift $c = 0.06$, (Table 7), for both asymptotic and exact probability measures and for all other means shifts, testing revealed robust rates for the data sampled from all of the normal type distributions (digit preference, discrete mass zero, normal, smooth symmetric, and uniform). This robust rejection rate pattern was also demonstrated at most small and small/unequal sample sizes, unlike Mood-Westenberg. Similar to the Mood-Westenberg, as the means shifted, non-robust results were detected for the data sampled from most non-normal type distributions (including asymmetric growth, discrete mass zero with gap, extreme asymmetric decay); however, unlike Mood-Westenberg, all indicators of these non-robust measures were in the conservative direction except the liberal rates found with the test results from asymmetric growth.

A particularly strong and unique outcome for Siegel-Tukey was noted for the non-normal extreme bimodal data set. At sample size (90, 90), Siegel-Tukey, unlike Mood-Westenberg, demonstrated robust measures at virtually all means

shifts for extreme bimodal (slight liberal exceptions were noted at 0.5% α level when means shift was at c = 0.02, 0.03, and 0.1). This strong robust rejection pattern for all means shifts was also noted in the data sampled from the extreme bimodal data for all equal sample sizes and for unequal sample sizes when α was 0.05.

Results demonstrated that the data sampled from the multi-modal lumpy data set was robust at lower means shifts but began to show conservative non-robust measures at means shifts generally at and above 0.09 for sample size (90, 90). However, many other sample sizes were robust at all means shifts. Results for data sampled from the exponential distribution became conservatively non-robust at means shift of c = 0.03 at sample size (90, 90). This was a general pattern for other large and equal sample sizes, although some smaller and unequal sample sizes maintained robust rates at higher mean shifts.

Siegel-Tukey's conservative non-robust rate exceptions, for non-normal distributions, multi-modal lumpy, and exponential, were deemed positive outcomes because this condition would obviate large pronouncements of nonsense variance changes. It did not demonstrate sample size instability that seemed pervasive throughout the study for Mood-Westenberg. At this point, after demonstrating large liberal rejection rates as the means shifted slightly with the non-normal type distributions, the Mood-Westenberg necessarily dropped out of consideration as a method to detect variance changes with respect to these distributions/data sets (though it maintained viability for exponential distributions and multi-modal lumpy data sets at lower means shift levels); however with the exception of the asymmetric growth data set, which measured liberal rejection rates, Siegel-Tukey demonstrated robust and conservatively robust rejection rates and thus continued as a viable instrument to detect heteroscedasticity for all other distributions/data sets provided power could be demonstrated next as the variance began to change.

Type-II Rejection Rates: For All Distributions/Data Sets, Variances are Unequal

During the final phase of the primary study, as assumptions were violated and variance changes simulated, the investigation focused upon reporting Mood-Westenberg and Siegel-Tukey asymptotic probabilities (Chi-squared and *Z*-scores, respectively) with nominal α of 0.05 and 0.01. The expectation was that power levels of at least 40% would be generally demonstrated.

With respect to the normal type distributions, both statistics generally demonstrated at least 40% power for all means shifts and variance changes for

large samples sizes (i.e., (30, 30) and (30, 90)), especially for $\alpha = 0.05$. Power (at sample size (30, 30) and above) approached 40% generally around variance change with $K = 1.75$-2 for α 0.05 and 0.01. For these normal type distributions, Siegel-Tukey typically demonstrated 40% power starting at smaller sample sizes (sample size (20, 20); Table 8) and often at lower levels of K changes ($K = 1.5$; Table 9) as compared to Mood-Westenberg (see also sample size (20, 20), uniform, for Siegel-Tukey's superior power; Table 10). Power for each statistic was shown to increase as α, variance, and sample size increased as demonstrated when the uniform sample size increased from (20, 20) (Table 10) to (45, 45) (Table 11) to (65, 65) (Table 12). While there were power improvements for both statistics as these parameters increased, Siegel-Tukey always demonstrated greater (or equal) power as compared to Mood-Westenberg at each point of comparison, sometimes yielding 20-40% more power at lower variance change levels.

For data sampled from non-normal distributions, both statistics reported much larger rejection rates as compared to the normal types when the variance changed and means shifted. This high rejection rate, starting from the smallest constant $K = 1.25$-3.5 (0.25), is reported for the representative data set, discrete mass zero with gap at sample sizes (45, 45) (Table 13). However, these large power rate results for the data sampled from non-normal distributions under Mood-Westenberg were meaningless due to the large liberal rejection rates noted for these when the variances were equal at $K = 1$ (see also large rate rejections 0.991-1 for discrete mass zero with gap and asymmetric decay in Table 7, at sample size (90, 90) when variances were equal).

However, given the conservative Type-I rejection rates (0.000) demonstrated when variances were equal for Siegel-Tukey, the large power it reported as variances changed is meaningful and impressive. For both small (e.g., (10, 10); Table 14) and large (e.g., (45, 45); Table 13) sample sizes, the Siegel-Tukey results for non-normal distributions, with the exception of asymmetric growth with many liberal Type-I rejection rates, had significant power that quickly approaching 99% at even the lowest levels of variance change (see also extreme bimodal; Table 15). For these non-normal power rates, a desired more gradual increase in power for Siegel-Tukey might have been demonstrated at lower levels of variance change between $K = 1$ and 1.25, but these levels were not tested. An impressive power finding was noted for the extreme bimodal data set under the Siegel-Tukey statistic, wherein the Type-I rejection rates were generally robust (instead of conservatively non-robust as Siegel-Tukey demonstrated with other non-normal distributions), particularly when sample sizes were equal (Table

7) and for unequal samples sizes when $\alpha = 0.05$. These robust findings, together with the high power noted in Table 15, renders the Siegel-Tukey test particularly useful in research settings where extreme bimodal data sets are common.

Finally, the results for both statistics with the data sampled from multi-modal lumpy and exponential demonstrated at least 40% power with large sample sizes (generally (30, 30), and above, including (30, 90)), especially when $\alpha = 0.05$. For Mood-Westenberg these results were attained typically at $K = 1.5$; for Siegel-Tukey at the lower $K = 1.25$. For the multi-modal lumpy data set with $\alpha = 0.05$ and the smallest variance change $K = 1.25$, 40% power was generally attained when sample size was (65, 65) for Mood-Westenberg and (30, 30) for Siegel-Tukey (Table 16, 17). For the exponential distribution (Table 18, 19), when $\alpha = 0.05$, 40% power was generally attained when $K = 1.5$ at sample size (30, 30) and (20, 20), respectively. Once again, Siegel-Tukey demonstrated greater or equal power at all comparison points than Mood-Westenberg for both of these distributions/data sets. For Mood-Westenberg, stable power was generally best when means shifts were below $c = 0.09$ for multi-modal lumpy and $c = 0.06$ for exponential due to some liberal non-robust Type-I rates at larger means shift levels. Siegel-Tukey was most powerful for these with lower means shifts ($c = 0.01$-0.08 for multi-modal lumpy and $c = 0.01$-0.03 for exponential) due to some conservative non-robust null rejections at larger mean shift levels.

Conclusion

Methods for Behrens-Fisher detection have been overlooked in statistical literature and, up to now, there have been no early warning or detective systems indicating the Behrens-Fisher condition exists. Siegel-Tukey appears promising as a method that might fill this void. Invoking the Siegel-Tukey statistic for the purpose of detecting variance changes could provide an effective precursor to the discovery of small yet important treatment effects in many research settings approaching Behrens-Fisher.

The Mood-Westenberg statistic also identified variance changes accompanied by slight mean shifts for normal type distributions, particularly with large sample sizes at or above $n = 30, 30$ (and at some smaller mean shifts for the multi-modal lumpy data set and the exponential distribution). However, Mood-Westenberg could not approach the levels of superior power demonstrated by Siegel-Tukey with these data sets/distributions and could not consistently demonstrate Siegel-Tukey's robust Type-I rejection rates at small sample sizes, especially when α was at 0.01.

Another significant comparative advantage demonstrated by the Siegel-Tukey statistic was its robust (or conservatively non-robust) and powerful results for non-normal distributions while Mood-Westenberg could not withstand the same means shift assumption violations for these types, demonstrating large liberal Type-I rejection rates. Therefore, as a detection tool for determining outcomes approaching Behrens-Fisher, the Mood-Westenberg statistic would be limited to research settings utilizing only normal type data distributions (best with larger sample sizes), the multi-modal lumpy data set, and the exponential distribution. Additionally, it is believed that the inability to stabilize Type-I rejection rates to approach nominal α level as sample sizes increased would render the Mood-Westenberg statistic generally less reliable in research settings.

Therefore, the Siegel-Tukey statistic might reasonably be promoted as the current statistic of choice in many scientific, educational and psychological research environments to detect heteroscedasticity whenever conditions approaching Behrens-Fisher arise with the concomitant problem of determining the existence of small means shift around zero. Siegel-Tukey demonstrated particularly strong measures for the extreme bimodal data set, often found within educational settings, when samples sizes were equal (or unequal at $\alpha = 0.05$). Siegel-Tukey's robust and powerful measures in detecting variance changes with all but one (asymmetric growth) of the 11 tested distributions/data sets demonstrated that it could be an important new instrument in the researcher's repertoire for data analysis. It has the potential to operate within a broad range of testing conditions to alert the researcher to the necessity of choosing an appropriate test statistic which could ultimately lead to the discovery of small treatments that might otherwise go unnoticed. The Siegel-Tukey statistic demonstrated its ability to be an effective precursor that would make known the need to replace testing statistics dependent on the equal variance assumptions, such as Student's-*t*, and to choose instead to apply any of the myriad of approximate Behrens-Fisher solutions, such as the Yuen's solution.

References

Behrens, W. V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher, 68*, 807-837.

Blair, R. C. (1987). Rangen (Version 1.0) [Software]. Boca Raton, FL: IBM.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16*(1), 129-132. doi: 10.2307/1267501

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dinneen, L. C., & Blakesley, B. C. (1973). Algorithm AS 62: A generator for the sampling distribution of the Mann-Whitney *U* statistic. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 22*(2), 269-273. doi: 10.2307/2346934

Fahoome, G. F. (2002). JMASM1: *RANGEN* 2.0 (*Fortran* 90/95). *Journal of Modern Applied Statistical Methods, 1*(1), 182-190. doi: 10.22237/jmasm/1020255960

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement, 34*(4), 789-799. doi: 10.1177/001316447403400406

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics, 6*(4), 391-398. doi: 10.1111/j.1469-1809.1935.tb02120.x

Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Annals of Eugenics, 9*(2), 174-180. doi: 10.1111/j.1469-1809.1939.tb02205.x

Hartley, H. O. (1950). The maximum *F*-ratio as a short-cut test for heterogeneity of variance. *Biometrika, 37*(3/4), 308-312. doi: 10.2307/2332383

Keppel, G. & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson Prentice Hall.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some non-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology, 30*(2), 213-221. doi: 10.1111/j.2044-8317.1977.tb00742.x

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583-621. doi: 10.2307/2280779

McSweeney, M., & Penfield, D. (1969). The normal scores test for the *c*-sample problem. *British Journal of Mathematical and Statistical Psychology, 22*(2), 177-192. doi: 10.1111/j.2044-8317.1969.tb00429.x

Micceri, T. (1986, November). *A futile search for the statistical chimera of normality*. Paper presented at the 31st Annual Convention of the Florida Educational Research Association, Tampa, FL.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037//0033-2909.105.1.156

Mood, A. M. (1950). *Introduction to the theory of statistics*. New York, NY: McGraw Hill.

Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests*. London, UK: Unwin Hyman Ltd.

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal, 14*(4), 493-498. doi: 10.2307/1162346

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods, 1*(2), 461-472. doi: 10.22237/jmasm/1036109940

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8*(2), 597-599.

Sawilowsky, S.S., & Blair, R.C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*(2), 353-360. doi: 10.1037/0033-2909.111.2.352

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and educational data sets. *Psychometrika, 55*(4), 729.

Sawilowsky, S. S., & Fahoome, G. F. (2003). *Statistics through Monte Carlo simulation with Fortran*. Oak Park, MI: JMASM.

Siegel, S. &. Tukey, J. W. (1960). A non-parametric sum of ranks procedure for relative spread in unpaired samples. *Journal of the American Statistical Association, 55*(291), 429-445. doi: 10.2307/2281906

Tomarken, A. J., & Serlin R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(1), 90-99. doi: 10.1037/0033-2909.99.1.90

Westenberg, J. (1948). Significance test for median and interquartile range in samples from continuous population of any form. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, 51*, 252-261.

Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. *Journal of Educational Statistics, 14*(3), 269-278. doi: 10.2307/1165019

Wilcox, R. R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation. *British Journal of Mathematical and Statistical Psychology, 48*(1), 99-114. doi: 10.1111/j.2044-8317.1995.tb01052.x

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilcox, R. R., Charlin, V., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA *F, W,* and *F** statistics. *Communications in Statistics – Simulation and Computation, 15*(4), 933-944. doi: 10.1080/03610918608812553

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83. doi: 10.2307/3001968

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika, 61*(1), 165-170. doi: 10.2307/2334299

Appendix A: Figures

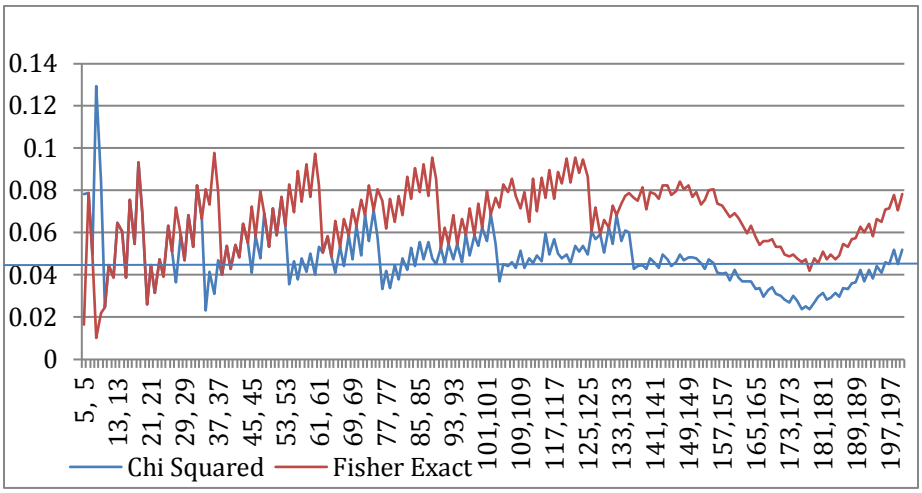


Figure 1. Mood-Westenberg Type-I error rate, comparisons between Chi Squared (blue) and Fisher Exact (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.05 α, 10,000 repetitions

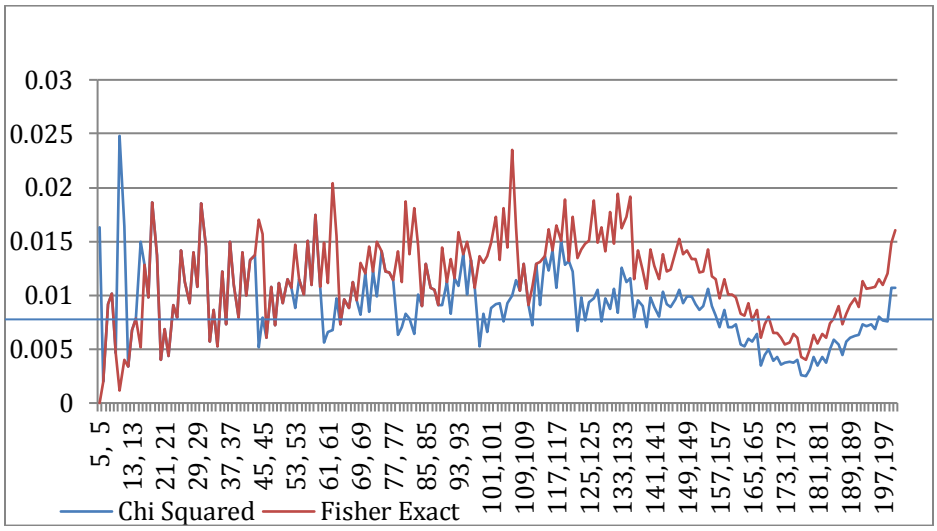


Figure 2. Mood-Westenberg Type-I error rate, comparisons between Chi Squared (blue) and Fisher Exact (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.01 α, 10,000 repetitions

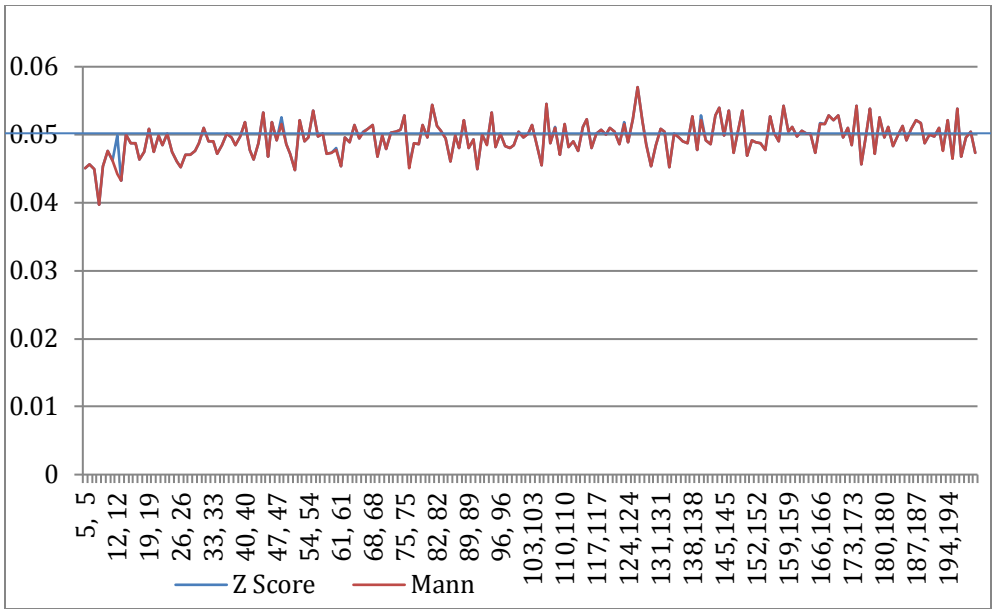


Figure 3. Siegel-Tukey Type-I error rate, comparisons between *Z* Scores (blue) and Mann-Whitney (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.05 *α*, 10,000 repetitions

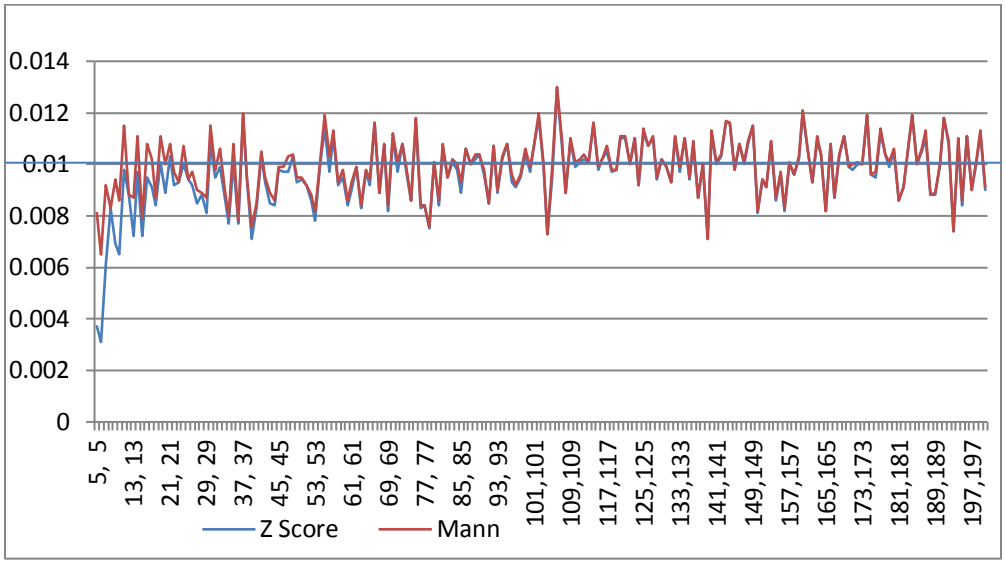


Figure 4. Siegel-Tukey Type-I error rate, comparisons between *Z* Scores (blue) and Mann-Whitney (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.01 *α*, 10,000 repetitions

Appendix B: Tables

Table 1. Type-I error rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional rest, for various sample sizes and *α* levels when sampling is from the normal distribution, 100,000 repetitions, variances are equal and means are equal

Mood-Westenberg								
	α							
	0.050		0.025		0.010		0.005	
Sample Size	A	E	A	E	A	E	A	E
5, 5	0.080	0.016	0.016	0.016	0.016	0.000	0.016	0.000
5, 15	0.033	0.033	0.033	0.033	0.033	0.000	0.000	0.000
10, 10	0.022	0.022	0.022	0.022	0.022	0.001	0.001	0.001
10, 30	0.066	0.066	0.008	0.008	0.008	0.008	0.008	0.008
15, 45	0.072	0.072	0.016	0.016	0.016	0.016	0.002	0.002
20, 20	0.026	0.026	0.026	0.026	0.004	0.004	0.004	0.004
30, 30	0.068	0.068	0.019	0.019	0.019	0.019	0.004	0.004
30, 90	0.056	0.056	0.020	0.020	0.006	0.020	0.006	0.006
45, 45	0.043	0.070	0.025	0.025	0.007	0.014	0.004	0.004
65, 65	0.041	0.063	0.026	0.026	0.010	0.010	0.006	0.006
90, 90	0.052	0.052	0.025	0.025	0.011	0.011	0.004	0.004
Siegel-Tukey								
5, 5	0.047	0.047	0.016	0.016	0.004	0.008	0.000	0.004
5, 15	0.058	0.048	0.025	0.021	0.010	0.010	0.004	0.004
10, 10	0.044	0.044	0.021	0.021	0.007	0.009	0.003	0.004
10, 30	0.051	0.047	0.024	0.024	0.010	0.010	0.004	0.004
15, 45	0.051	0.050	0.027	0.025	0.010	0.010	0.005	0.005
20, 20	0.048	0.048	0.025	0.025	0.010	0.010	0.004	0.005
30, 30	0.050	0.050	0.023	0.024	0.009	0.010	0.005	0.005
30, 90	0.050	0.049	0.025	0.024	0.009	0.010	0.005	0.005
45, 45	0.049	0.049	0.024	0.024	0.010	0.010	0.005	0.005
65, 65	0.049	0.049	0.024	0.024	0.010	0.010	0.005	0.005
90, 90	0.050	0.050	0.025	0.025	0.010	0.010	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 2. Type-I error rate averages for all sample sizes (5, 5) to (200, 200) for 10,000 repetitions, Normal distribution

Mood-Westenberg								
α								
0.050		0.025		0.010		0.005		
A	E	A	E	A	E	A	E	
0.048	0.067	0.024	0.031	0.009	0.012	0.005	0.005	

Table 2, continued.

Siegel-Tukey							
α							
0.050		0.025		0.010		0.005	
A	E	A	E	A	E	A	E
0.049	0.049	0.024	0.025	0.010	0.010	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 3. Type-I error rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for sample size (45, 45) and *α* levels when sampling is from all distributions/data sets, 100,000 repetitions, variances are equal, and means are equal

Mood-Westenberg								
	α							
	0.050		0.025		0.010		0.005	
Distribution	A	E	A	E	A	E	A	E
Asym Growth	0.040	0.067	0.024	0.024	0.007	0.013	0.003	0.003
Digit pref	0.042	0.069	0.024	0.024	0.007	0.014	0.004	0.004
Disc mass zero	0.040	0.066	0.023	0.023	0.007	0.012	0.003	0.003
Disc mass zero gap	0.004	0.008	0.002	0.002	0.000	0.001	0.000	0.000
Exponential	0.043	0.071	0.025	0.025	0.007	0.014	0.004	0.004
Extrm asym decay	0.021	0.039	0.011	0.011	0.002	0.005	0.001	0.001
Extrm bimodal	0.022	0.041	0.011	0.011	0.002	0.005	0.001	0.001
Multi-modal lumpy	0.042	0.069	0.024	0.024	0.007	0.014	0.004	0.004
Normal	0.043	0.070	0.025	0.025	0.007	0.014	0.004	0.004
Smooth sym	0.040	0.066	0.023	0.023	0.007	0.013	0.003	0.003
Uni	0.043	0.070	0.025	0.025	0.008	0.015	0.004	0.004
Siegel-Tukey								
Asym Growth	0.046	0.047	0.022	0.022	0.008	0.009	0.004	0.004
Digit pref	0.049	0.050	0.024	0.025	0.009	0.010	0.005	0.005
Disc mass zero	0.047	0.048	0.023	0.024	0.009	0.009	0.004	0.005
Disc mass zero gap	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
Exponential	0.050	0.050	0.026	0.026	0.010	0.010	0.005	0.005
Extrm asym decay	0.011	0.011	0.003	0.003	0.001	0.001	0.000	0.000
Extrm bimodal	0.023	0.024	0.009	0.009	0.003	0.003	0.001	0.001
Multi-modal lumpy	0.049	0.050	0.024	0.025	0.009	0.010	0.005	0.005
Normal	0.049	0.049	0.024	0.024	0.010	0.010	0.005	0.005
Smooth sym	0.048	0.048	0.023	0.024	0.009	0.009	0.004	0.004
Uni	0.049	0.049	0.025	0.025	0.009	0.009	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 4. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various *α* levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, means are equal, and variance change is 1.25

	Mood-Westenberg							
	α							
	0.050		**0.025**		**0.010**		**0.005**	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Asym Growth	0.457	0.457	0.369	0.369	0.289	0.289	0.219	0.219
Digit pref	0.265	0.265	0.179	0.179	0.114	0.114	0.068	0.068
Disc mass zero	0.197	0.197	0.128	0.128	0.078	0.078	0.044	0.044
Disc mass zero gap			0.999	0.999	0.996	0.996	0.991	0.991
Exponential	0.478	0.478	0.360	0.360	0.256	0.256	0.170	0.170
Extrm asym decay					0.999	0.999	0.999	0.999
Extrm bimodal	0.897	0.897	0.852	0.852	0.795	0.795	0.726	0.726
Multi-modal lumpy	0.668	0.668	0.559	0.559	0.446	0.446	0.334	0.334
Normal	0.257	0.257	0.169	0.169	0.102	0.102	0.058	0.058
Smooth sym	0.165	0.165	0.104	0.104	0.061	0.061	0.034	0.034
Uni	0.330	0.330	0.230	0.230	0.150	0.150	0.090	0.090
	Siegel-Tukey							
Asym Growth	0.886	0.886	0.815	0.816	0.703	0.706	0.614	0.616
Digit pref	0.512	0.513	0.389	0.389	0.258	0.261	0.184	0.186
Disc mass zero	0.568	0.569	0.446	0.447	0.308	0.310	0.225	0.227
Disc mass zero gap								
Exponential	0.830	0.830	0.735	0.735	0.603	0.605	0.502	0.504
Extrm asym decay					0.999	0.999	0.999	0.999
Extrm bimodal							0.999	0.999
Multi-modal lumpy	0.846	0.846	0.758	0.758	0.630	0.632	0.531	0.533
Normal	0.495	0.495	0.370	0.370	0.240	0.242	0.169	0.170
Smooth sym	0.550	0.550	0.425	0.426	0.288	0.290	0.210	0.212
Uni	0.750	0.750	0.639	0.639	0.494	0.496	0.394	0.397

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 5. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various *α* levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, means are equal, and variance change is 1.5

Mood-Westenberg

	α							
	0.050		**0.025**		**0.010**		**0.005**	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Asym Growth	0.888	0.888	0.827	0.827	0.746	0.746	0.651	0.651
Digit pref	0.570	0.570	0.458	0.458	0.349	0.349	0.250	0.250
Disc mass zero	0.615	0.615	0.515	0.515	0.416	0.416	0.322	0.322
Disc mass zero gap			0.999	0.999	0.997	0.997	0.991	0.991
Exponential	0.916	0.916	0.861	0.861	0.787	0.787	0.692	0.692
Extrm asym decay								
Extrm bimodal	0.897	0.897	0.851	0.851	0.794	0.794	0.726	0.726
Multi-modal lumpy	0.971	0.971	0.946	0.946	0.906	0.906	0.849	0.849
Normal	0.643	0.643	0.527	0.527	0.407	0.407	0.293	0.293
Smooth sym	0.651	0.651	0.543	0.543	0.433	0.433	0.328	0.328
Uni	0.776	0.776	0.678	0.678	0.567	0.567	0.449	0.449

Siegel-Tukey

Asym Growth	0.997	0.997	0.994	0.994	0.983	0.983	0.969	0.970
Digit pref	0.896	0.896	0.829	0.830	0.720	0.722	0.630	0.633
Disc mass zero	0.894	0.894	0.826	0.826	0.715	0.717	0.625	0.628
Disc mass zero gap								
Exponential	0.995	0.995	0.988	0.988	0.970	0.970	0.948	0.949
Extrm asym decay								
Extrm bimodal							0.999	0.999
Multi-modal lumpy	0.998	0.998	0.996	0.996	0.987	0.988	0.977	0.978
Normal	0.899	0.899	0.831	0.831	0.721	0.722	0.629	0.631
Smooth sym	0.902	0.902	0.835	0.836	0.729	0.732	0.641	0.644
Uni	0.988	0.988	0.974	0.974	0.942	0.943	0.907	0.908

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 6. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various *α* levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, means are equal, and variance change is 2.25

Mood-Westenberg

	α							
	0.050		**0.025**		**0.010**		**0.005**	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Asym Growth								
Digit pref	0.985	0.985	0.971	0.971	0.948	0.948	0.913	0.913
Disc mass zero	0.990	0.990	0.981	0.981	0.965	0.965	0.940	0.940
Disc mass zero gap			0.999	0.999	0.996	0.996	0.990	0.990
Exponential							0.999	0.999
Extrm asym decay								
Extrm bimodal								
Multi-modal lumpy								
Normal	0.995	0.995	0.988	0.988	0.976	0.976	0.953	0.953
Smooth sym	0.985	0.985	0.970	0.970	0.946	0.946	0.909	0.909
Uni	0.999	0.999	0.998	0.998	0.996	0.996	0.990	0.990

Siegel-Tukey

Asym Growth								
Digit pref					0.999	0.999	0.997	0.997
Disc mass zero					0.999	0.999	0.999	0.999
Disc mass zero gap								
Exponential								
Extrm asym decay								
Extrm bimodal								
Multi-modal lumpy								
Normal					0.999	0.999	0.999	0.999
Smooth sym							0.999	0.999
Uni								

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 7. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various *α* levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, variances are equal, and means shift is 0.06

Mood-Westenberg

	α							
	0.050		0.025		0.010		0.005	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Asym Growth	0.240	0.240	0.163	0.163	0.105	0.105	0.063	0.063
Digit pref	0.063	0.063	0.031	0.031	0.014	0.014	0.006	0.006
Disc mass zero	0.073	0.073	0.039	0.039	0.019	0.019	0.009	0.009
Disc mass zero gap			0.999	0.999	0.996	0.996	0.991	0.991
Exponential	0.071	0.071	0.037	0.037	0.018	0.018	0.008	0.008
Extrm asym decay			0.999	0.999	0.998	0.998	0.997	0.997
Extrm bimodal	0.537	0.537	0.459	0.459	0.383	0.383	0.310	0.310
Multi-modal lumpy	0.060	0.060	0.030	0.030	0.014	0.014	0.006	0.006
Normal	0.053	0.053	0.025	0.025	0.011	0.011	0.005	0.005
Smooth sym	0.065	0.065	0.033	0.033	0.015	0.015	0.007	0.007
Uni	0.052	0.052	0.025	0.025	0.010	0.010	0.004	0.004

Siegel-Tukey

Asym Growth	0.298	0.298	0.198	0.198	0.111	0.112	0.071	0.072
Digit pref	0.050	0.050	0.025	0.026	0.010	0.011	0.005	0.005
Disc mass zero	0.040	0.040	0.020	0.020	0.008	0.008	0.004	0.004
Disc mass zero gap	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Exponential	0.011	0.011	0.005	0.005	0.001	0.001	0.001	0.001
Extrm asym decay	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Extrm bimodal	0.056	0.056	0.031	0.031	0.014	0.014	0.007	0.007
Multi-modal lumpy	0.038	0.038	0.018	0.018	0.007	0.007	0.003	0.003
Normal	0.050	0.050	0.025	0.025	0.010	0.010	0.005	0.005
Smooth sym	0.050	0.050	0.025	0.025	0.010	0.010	0.005	0.005
Uni	0.048	0.048	0.024	0.024	0.010	0.010	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 8. Power rates for one-tailed directional test for digit preference data set, various means shifts and variance changes for sample size (20, 20), 100,000 repetitions, $\alpha = 0.05$

	Mood-Westenberg Chi-squared										
Means	**Variance change**										
shift	**1.00**	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.024	0.053	0.119	0.195	0.307	0.416	0.495	0.578	0.676	0.728	0.768
0.01	0.028	0.054	0.117	0.205	0.307	0.414	0.495	0.579	0.675	0.727	0.766
0.02	0.027	0.049	0.119	0.203	0.305	0.413	0.499	0.578	0.676	0.731	0.766
0.03	0.027	0.051	0.113	0.205	0.307	0.415	0.501	0.570	0.675	0.732	0.768
0.04	0.026	0.051	0.114	0.202	0.307	0.408	0.504	0.569	0.676	0.727	0.769
0.05	0.027	0.055	0.112	0.201	0.306	0.408	0.505	0.568	0.675	0.728	0.769
0.06	0.027	0.055	0.112	0.199	0.302	0.409	0.503	0.620	0.676	0.722	0.766
0.07	0.026	0.055	0.112	0.200	0.301	0.402	0.501	0.620	0.674	0.726	0.773
0.08	0.027	0.057	0.113	0.196	0.302	0.401	0.499	0.620	0.674	0.724	0.773
0.09	0.027	0.057	0.115	0.197	0.301	0.404	0.499	0.621	0.674	0.720	0.771
0.10	0.027	0.057	0.117	0.198	0.301	0.427	0.500	0.622	0.675	0.723	0.771
0.11	0.027	0.058	0.119	0.200	0.302	0.429	0.498	0.623	0.678	0.721	0.774
0.12	0.026	0.057	0.119	0.199	0.303	0.429	0.498	0.622	0.679	0.717	0.773
	Siegel-Tukey *Z*-score										
0.00	0.048	0.177	0.366	0.535	0.687	0.789	0.849	0.897	0.933	0.954	0.963
0.01	0.050	0.179	0.362	0.543	0.687	0.788	0.849	0.897	0.932	0.953	0.963
0.02	0.050	0.168	0.363	0.540	0.686	0.788	0.853	0.896	0.933	0.948	0.963
0.03	0.050	0.168	0.354	0.543	0.688	0.789	0.853	0.897	0.933	0.949	0.964
0.04	0.049	0.169	0.355	0.524	0.690	0.794	0.853	0.897	0.934	0.947	0.964
0.05	0.049	0.179	0.352	0.527	0.685	0.792	0.855	0.897	0.932	0.947	0.964
0.06	0.049	0.177	0.352	0.525	0.673	0.793	0.855	0.906	0.933	0.947	0.964
0.07	0.049	0.178	0.351	0.521	0.671	0.774	0.855	0.904	0.932	0.949	0.966
0.08	0.050	0.185	0.354	0.525	0.669	0.774	0.843	0.907	0.932	0.948	0.965
0.09	0.050	0.186	0.356	0.526	0.672	0.773	0.842	0.906	0.931	0.954	0.965
0.10	0.050	0.185	0.357	0.528	0.670	0.779	0.843	0.895	0.933	0.954	0.964
0.11	0.050	0.186	0.361	0.535	0.671	0.780	0.844	0.893	0.929	0.954	0.965
0.12	0.050	0.184	0.362	0.534	0.670	0.782	0.841	0.896	0.931	0.948	0.965

Table 9. Power rates for one-tailed directional test for digit preference data set, various means shifts and variance changes for sample size (30, 30), 100,000 repetitions, $\alpha = 0.05$

Means shift	Variance change										
Mood-Westenberg Chi-squared											
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.065	0.145	0.292	0.441	0.613	0.743	0.817	0.874	0.930	0.951	0.965
0.01	0.073	0.142	0.291	0.457	0.611	0.744	0.813	0.873	0.930	0.951	0.965
0.02	0.073	0.135	0.291	0.458	0.615	0.742	0.821	0.875	0.930	0.952	0.964
0.03	0.073	0.134	0.279	0.456	0.612	0.743	0.820	0.867	0.931	0.953	0.965
0.04	0.072	0.133	0.281	0.454	0.612	0.730	0.821	0.866	0.929	0.953	0.965
0.05	0.072	0.141	0.276	0.452	0.614	0.730	0.820	0.866	0.927	0.954	0.965
0.06	0.073	0.143	0.278	0.451	0.611	0.730	0.821	0.904	0.931	0.949	0.964
0.07	0.073	0.142	0.278	0.454	0.607	0.727	0.819	0.905	0.930	0.950	0.966
0.08	0.073	0.151	0.281	0.445	0.611	0.727	0.818	0.902	0.930	0.949	0.967
0.09	0.075	0.150	0.280	0.444	0.613	0.727	0.819	0.903	0.931	0.948	0.967
0.10	0.074	0.150	0.292	0.443	0.610	0.761	0.819	0.908	0.930	0.948	0.967
0.11	0.073	0.148	0.292	0.447	0.611	0.760	0.818	0.907	0.933	0.948	0.967
0.12	0.073	0.153	0.292	0.443	0.610	0.762	0.818	0.908	0.932	0.948	0.967
Siegel-Tukey *Z*-score											
0.00	0.047	0.237	0.498	0.706	0.849	0.922	0.957	0.977	0.989	0.994	0.996
0.01	0.051	0.234	0.498	0.711	0.849	0.922	0.957	0.978	0.989	0.994	0.996
0.02	0.051	0.219	0.498	0.714	0.849	0.922	0.958	0.976	0.988	0.992	0.996
0.03	0.053	0.217	0.482	0.710	0.849	0.922	0.960	0.978	0.989	0.993	0.996
0.04	0.051	0.218	0.482	0.690	0.849	0.926	0.959	0.976	0.989	0.992	0.996
0.05	0.051	0.230	0.479	0.690	0.849	0.925	0.958	0.977	0.988	0.993	0.996
0.06	0.050	0.229	0.482	0.692	0.835	0.926	0.959	0.980	0.989	0.993	0.996
0.07	0.052	0.232	0.479	0.691	0.832	0.913	0.959	0.980	0.988	0.992	0.997
0.08	0.052	0.247	0.483	0.694	0.834	0.912	0.952	0.980	0.988	0.992	0.996
0.09	0.051	0.244	0.481	0.695	0.835	0.912	0.952	0.980	0.988	0.994	0.997
0.10	0.052	0.246	0.489	0.691	0.835	0.915	0.951	0.975	0.989	0.994	0.996
0.11	0.053	0.245	0.488	0.703	0.834	0.915	0.951	0.975	0.989	0.994	0.997
0.12	0.051	0.242	0.490	0.699	0.833	0.915	0.951	0.976	0.988	0.992	0.996

Table 10. Power rates for one-tailed directional test for uniform distribution, various means shifts and variance changes for sample size (20, 20), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.026	0.067	0.179	0.331	0.480	0.610	0.714	0.791	0.846	0.885	0.914
0.01	0.024	0.068	0.182	0.330	0.484	0.610	0.715	0.790	0.845	0.887	0.915
0.02	0.025	0.069	0.180	0.333	0.484	0.610	0.712	0.789	0.846	0.884	0.914
0.03	0.025	0.067	0.181	0.330	0.485	0.613	0.714	0.790	0.845	0.884	0.913
0.04	0.026	0.068	0.180	0.331	0.484	0.612	0.715	0.791	0.846	0.885	0.913
0.05	0.026	0.068	0.180	0.331	0.481	0.609	0.712	0.791	0.845	0.885	0.915
0.06	0.026	0.067	0.182	0.330	0.482	0.613	0.712	0.791	0.843	0.886	0.914
0.07	0.025	0.069	0.179	0.331	0.481	0.612	0.717	0.791	0.845	0.885	0.914
0.08	0.026	0.068	0.182	0.330	0.483	0.611	0.714	0.790	0.846	0.886	0.914
0.09	0.026	0.069	0.179	0.329	0.482	0.611	0.713	0.790	0.844	0.883	0.914
0.10	0.026	0.069	0.178	0.332	0.482	0.612	0.711	0.789	0.844	0.885	0.914
0.11	0.026	0.068	0.182	0.332	0.484	0.613	0.714	0.789	0.844	0.887	0.914
0.12	0.025	0.068	0.179	0.332	0.481	0.611	0.715	0.788	0.844	0.884	0.916

Siegel-Tukey *Z*-score

0.00	0.048	0.272	0.548	0.745	0.859	0.922	0.955	0.973	0.984	0.989	0.994
0.01	0.046	0.272	0.548	0.744	0.860	0.922	0.955	0.973	0.984	0.990	0.994
0.02	0.048	0.273	0.548	0.746	0.861	0.921	0.955	0.974	0.984	0.989	0.993
0.03	0.047	0.269	0.549	0.745	0.861	0.922	0.955	0.974	0.985	0.990	0.993
0.04	0.048	0.272	0.547	0.746	0.860	0.921	0.956	0.974	0.984	0.990	0.993
0.05	0.048	0.272	0.549	0.745	0.859	0.922	0.955	0.973	0.984	0.990	0.994
0.06	0.049	0.270	0.547	0.743	0.858	0.922	0.956	0.974	0.984	0.990	0.993
0.07	0.048	0.269	0.545	0.745	0.860	0.923	0.955	0.974	0.985	0.990	0.993
0.08	0.047	0.273	0.547	0.745	0.859	0.920	0.955	0.974	0.983	0.990	0.993
0.09	0.048	0.271	0.546	0.743	0.859	0.921	0.955	0.973	0.983	0.990	0.994
0.10	0.046	0.269	0.545	0.745	0.859	0.922	0.954	0.973	0.983	0.990	0.993
0.11	0.047	0.266	0.545	0.743	0.859	0.922	0.956	0.974	0.984	0.990	0.993
0.12	0.047	0.267	0.545	0.744	0.857	0.923	0.955	0.973	0.983	0.990	0.994

Table 11. Power rates for one-tailed directional test for uniform distribution, various means shifts and variance changes for sample size (45, 45), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.043	0.174	0.468	0.730	0.882	0.952	0.982	0.993	0.997	0.999	
0.01	0.044	0.176	0.468	0.730	0.882	0.953	0.983	0.993	0.997	0.999	0.999
0.02	0.044	0.175	0.468	0.731	0.884	0.952	0.981	0.993	0.997	0.999	0.999
0.03	0.044	0.174	0.465	0.731	0.883	0.953	0.981	0.993	0.997	0.999	
0.04	0.043	0.174	0.470	0.733	0.881	0.952	0.981	0.993	0.997	0.999	0.999
0.05	0.044	0.172	0.469	0.731	0.882	0.952	0.981	0.993	0.997	0.999	0.999
0.06	0.044	0.173	0.468	0.731	0.883	0.953	0.981	0.993	0.997	0.999	
0.07	0.043	0.175	0.465	0.731	0.883	0.953	0.981	0.993	0.997	0.999	
0.08	0.044	0.176	0.467	0.732	0.883	0.953	0.982	0.993	0.997	0.999	
0.09	0.042	0.174	0.469	0.730	0.883	0.952	0.981	0.992	0.997	0.999	
0.10	0.044	0.174	0.467	0.732	0.883	0.952	0.981	0.993	0.997	0.999	0.999
0.11	0.044	0.175	0.468	0.730	0.882	0.953	0.981	0.993	0.997	0.999	0.999
0.12	0.045	0.171	0.466	0.729	0.881	0.953	0.982	0.993	0.997	0.999	

Siegel-Tukey *Z*-score

0.00	0.049	0.493	0.865	0.972	0.995	0.999
0.01	0.050	0.493	0.863	0.973	0.995	0.999
0.02	0.050	0.492	0.865	0.972	0.995	0.999
0.03	0.050	0.492	0.862	0.973	0.995	0.999
0.04	0.050	0.493	0.864	0.973	0.994	0.999
0.05	0.050	0.491	0.866	0.972	0.995	0.999
0.06	0.050	0.491	0.862	0.972	0.995	0.999
0.07	0.049	0.491	0.862	0.972	0.994	0.999
0.08	0.050	0.491	0.863	0.972	0.995	0.999
0.09	0.048	0.489	0.863	0.973	0.995	0.999
0.10	0.050	0.488	0.862	0.971	0.995	0.999
0.11	0.049	0.491	0.862	0.973	0.995	0.999
0.12	0.049	0.486	0.861	0.972	0.995	0.999

Table 12. Power rates for one-tailed directional test for uniform distribution, various means shifts and variance changes for sample size (65, 65), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.043	0.230	0.609	0.867	0.964	0.992	0.998				
0.01	0.042	0.229	0.611	0.868	0.964	0.991	0.998	0.999			
0.02	0.043	0.230	0.612	0.865	0.964	0.991	0.998	0.999			
0.03	0.044	0.232	0.610	0.867	0.963	0.991	0.998	0.999			
0.04	0.042	0.230	0.612	0.869	0.964	0.991	0.998				
0.05	0.043	0.232	0.611	0.867	0.965	0.991	0.998				
0.06	0.042	0.232	0.610	0.867	0.964	0.991	0.998				
0.07	0.041	0.229	0.611	0.867	0.965	0.992	0.998	0.999			
0.08	0.043	0.229	0.613	0.868	0.965	0.991	0.998				
0.09	0.043	0.230	0.613	0.867	0.965	0.991	0.998				
0.10	0.042	0.232	0.613	0.866	0.964	0.992	0.998				
0.11	0.043	0.228	0.612	0.867	0.964	0.991	0.998				
0.12	0.041	0.229	0.611	0.867	0.965	0.992	0.998	0.999			

Siegel-Tukey *Z*-score

0.00	0.050	0.623	0.951	0.996
0.01	0.048	0.623	0.952	0.996
0.02	0.050	0.626	0.952	0.996
0.03	0.050	0.627	0.951	0.996
0.04	0.049	0.626	0.953	0.996
0.05	0.049	0.625	0.952	0.996
0.06	0.050	0.623	0.951	0.996
0.07	0.048	0.622	0.951	0.996
0.08	0.049	0.625	0.951	0.996
0.09	0.049	0.623	0.952	0.996
0.10	0.049	0.623	0.951	0.996
0.11	0.049	0.620	0.951	0.996
0.12	0.050	0.620	0.950	0.996

Table 13. Power rates for one-tailed directional test for discrete mass zero with gap data set, various means shifts and variance changes for sample size (45, 45), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.004	0.959	0.960	0.960	0.961	0.957	0.957	0.956	0.955	0.955	0.957
0.01	0.960	0.960	0.961	0.960	0.961	0.956	0.956	0.957	0.956	0.956	0.957
0.02	0.961	0.960	0.961	0.959	0.961	0.961	0.956	0.956	0.957	0.957	0.955
0.03	0.961	0.961	0.960	0.961	0.960	0.960	0.957	0.956	0.957	0.957	0.956
0.04	0.960	0.961	0.960	0.960	0.959	0.960	0.957	0.955	0.955	0.956	0.957
0.05	0.961	0.960	0.960	0.961	0.960	0.959	0.956	0.957	0.956	0.956	0.957
0.06	0.960	0.960	0.961	0.961	0.961	0.960	0.956	0.956	0.957	0.955	0.956
0.07	0.960	0.960	0.960	0.961	0.961	0.960	0.956	0.956	0.956	0.956	0.956
0.08	0.961	0.961	0.960	0.960	0.961	0.959	0.961	0.955	0.957	0.955	0.956
0.09	0.960	0.961	0.960	0.959	0.960	0.961	0.961	0.956	0.955	0.957	0.956
0.10	0.961	0.960	0.960	0.961	0.961	0.961	0.960	0.956	0.955	0.956	0.956
0.11	0.960	0.961	0.960	0.960	0.960	0.960	0.961	0.955	0.957	0.957	0.956
0.12	0.961	0.961	0.961	0.960	0.961	0.960	0.960	0.957	0.957	0.957	0.956

Siegel-Tukey *Z*-score

0.00	0.001	0.997	0.996	0.996	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.01	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.02	0.000	0.997	0.996	0.996	0.997	0.997	0.996	0.996	0.996	0.996	0.996
0.03	0.000	0.997	0.996	0.997	0.996	0.996	0.996	0.996	0.996	0.996	0.996
0.04	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.05	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.06	0.000	0.997	0.997	0.997	0.997	0.997	0.996	0.996	0.997	0.996	0.996
0.07	0.000	0.996	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996
0.08	0.000	0.997	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996
0.09	0.000	0.996	0.997	0.996	0.997	0.997	0.997	0.996	0.996	0.996	0.996
0.10	0.000	0.996	0.997	0.997	0.997	0.996	0.997	0.996	0.996	0.996	0.996
0.11	0.000	0.997	0.996	0.997	0.997	0.996	0.997	0.996	0.996	0.996	0.996
0.12	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996

Table 14. Power rates for one-tailed directional test for discrete mass zero with gap data set, various means shifts and variance changes for sample size (10, 10), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared											
Means shift	Variance change										
	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.005	0.310	0.316	0.314	0.315	0.308	0.307	0.309	0.308	0.307	0.308
0.01	0.248	0.308	0.314	0.315	0.314	0.310	0.309	0.308	0.308	0.310	0.305
0.02	0.247	0.310	0.316	0.318	0.314	0.313	0.307	0.308	0.308	0.308	0.309
0.03	0.249	0.309	0.313	0.315	0.316	0.315	0.309	0.311	0.311	0.311	0.310
0.04	0.246	0.310	0.316	0.316	0.315	0.316	0.309	0.309	0.309	0.311	0.310
0.05	0.246	0.310	0.317	0.312	0.314	0.315	0.308	0.310	0.308	0.309	0.307
0.06	0.248	0.311	0.315	0.317	0.312	0.316	0.310	0.308	0.306	0.306	0.309
0.07	0.246	0.313	0.316	0.317	0.315	0.313	0.308	0.309	0.305	0.309	0.309
0.08	0.245	0.311	0.314	0.314	0.317	0.314	0.315	0.309	0.306	0.306	0.308
0.09	0.249	0.312	0.315	0.314	0.315	0.312	0.315	0.308	0.309	0.309	0.311
0.10	0.244	0.313	0.316	0.315	0.316	0.315	0.315	0.310	0.308	0.311	0.309
0.11	0.247	0.311	0.315	0.314	0.314	0.317	0.313	0.307	0.311	0.310	0.311
0.12	0.247	0.308	0.314	0.315	0.314	0.315	0.314	0.310	0.312	0.310	0.308
Siegel-Tukey *Z*-score											
0.00	0.000	0.619	0.620	0.619	0.619	0.612	0.610	0.611	0.612	0.610	0.611
0.01	0.000	0.617	0.620	0.624	0.621	0.614	0.612	0.613	0.611	0.612	0.609
0.02	0.000	0.617	0.623	0.622	0.621	0.623	0.612	0.610	0.611	0.610	0.611
0.03	0.000	0.619	0.621	0.624	0.623	0.620	0.612	0.613	0.613	0.615	0.615
0.04	0.000	0.619	0.623	0.621	0.620	0.622	0.613	0.613	0.611	0.611	0.610
0.05	0.000	0.619	0.623	0.621	0.619	0.620	0.610	0.612	0.612	0.613	0.612
0.06	0.000	0.621	0.622	0.623	0.621	0.624	0.613	0.611	0.612	0.610	0.611
0.07	0.000	0.622	0.622	0.623	0.620	0.620	0.613	0.612	0.609	0.610	0.612
0.08	0.000	0.619	0.623	0.622	0.622	0.620	0.620	0.613	0.612	0.610	0.609
0.09	0.000	0.620	0.620	0.622	0.619	0.621	0.623	0.612	0.615	0.613	0.614
0.10	0.000	0.623	0.621	0.621	0.622	0.624	0.623	0.612	0.613	0.614	0.612
0.11	0.000	0.621	0.622	0.621	0.622	0.621	0.618	0.608	0.614	0.615	0.613
0.12	0.000	0.618	0.622	0.620	0.621	0.621	0.622	0.613	0.614	0.614	0.613

Table 15. Power rates for one-tailed directional test for discrete mass zero with gap data set, various means shifts and variance changes for sample size (10, 10), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.022	0.701	0.705	0.700							
0.01	0.347	0.699	0.701	0.701							
0.02	0.347	0.699	0.702	0.701							
0.03	0.349	0.703	0.701	0.704							
0.04	0.349	0.701	0.701	0.699							
0.05	0.349	0.700	0.701	0.701							
0.06	0.345	0.700	0.701	0.702							
0.07	0.346	0.701	0.703	0.703							
0.08	0.348	0.700	0.703	0.701							
0.09	0.347	0.702	0.702	0.700							
0.10	0.349	0.699	0.701	0.702							
0.11	0.350	0.702	0.702	0.702							
0.12	0.346	0.702	0.702	0.702							

Siegel-Tukey *Z*-score

0.00	0.023	0.991	0.991	0.992							
0.01	0.055	0.991	0.991	0.992							
0.02	0.054	0.991	0.992	0.992							
0.03	0.055	0.991	0.992	0.992							
0.04	0.054	0.991	0.991	0.992							
0.05	0.054	0.991	0.991	0.992							
0.06	0.055	0.991	0.991	0.992							
0.07	0.054	0.991	0.991	0.992							
0.08	0.054	0.991	0.991	0.992							
0.09	0.054	0.992	0.991	0.992							
0.10	0.053	0.992	0.991	0.992							
0.11	0.053	0.991	0.991	0.992							
0.12	0.052	0.992	0.991	0.992							

Table 16. Power rates for one-tailed directional test for multi-modal lumpy data set, various means shifts and variance changes for sample size (30, 30), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared											
Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.068	0.303	0.652	0.846	0.936	0.971	0.988	0.995	0.998	0.999	0.999
0.01	0.074	0.272	0.624	0.840	0.935	0.971	0.988	0.995	0.998	0.999	
0.02	0.072	0.273	0.623	0.841	0.924	0.969	0.988	0.995	0.998	0.999	0.999
0.03	0.072	0.266	0.623	0.840	0.923	0.970	0.988	0.995	0.998	0.999	0.999
0.04	0.073	0.266	0.625	0.823	0.922	0.969	0.988	0.995	0.998	0.999	0.999
0.05	0.073	0.261	0.590	0.823	0.925	0.968	0.988	0.994	0.998	0.999	0.999
0.06	0.074	0.263	0.591	0.817	0.923	0.967	0.987	0.994	0.998	0.999	0.999
0.07	0.071	0.258	0.590	0.818	0.925	0.968	0.985	0.994	0.997	0.999	0.999
0.08	0.074	0.258	0.590	0.817	0.924	0.968	0.985	0.994	0.997	0.998	0.999
0.09	0.080	0.247	0.592	0.814	0.923	0.968	0.985	0.994	0.998	0.999	0.999
0.10	0.078	0.249	0.587	0.805	0.914	0.966	0.985	0.994	0.998	0.999	0.999
0.11	0.079	0.221	0.589	0.804	0.915	0.966	0.984	0.993	0.997	0.999	0.999
0.12	0.077	0.221	0.586	0.798	0.914	0.965	0.984	0.994	0.997	0.999	0.999

Siegel-Tukey *Z*-score											
0.00	0.049	0.444	0.831	0.961	0.992	0.998	0.999				
0.01	0.043	0.430	0.812	0.956	0.992	0.998					
0.02	0.043	0.431	0.811	0.958	0.989	0.998	0.999				
0.03	0.043	0.418	0.812	0.957	0.989	0.998	0.999				
0.04	0.043	0.417	0.814	0.952	0.989	0.997	0.999				
0.05	0.044	0.399	0.788	0.953	0.989	0.997	0.999				
0.06	0.043	0.399	0.790	0.948	0.989	0.997	0.999				
0.07	0.042	0.388	0.789	0.949	0.989	0.997	0.999				
0.08	0.044	0.388	0.788	0.945	0.989	0.997	0.999				
0.09	0.031	0.376	0.792	0.943	0.989	0.997	0.999				
0.10	0.032	0.378	0.773	0.939	0.985	0.997	0.999				
0.11	0.032	0.357	0.774	0.940	0.985	0.997	0.999				
0.12	0.032	0.357	0.772	0.940	0.985	0.997	0.999				

Table 17. Power rates for one-tailed directional test for multi-modal lumpy data set, various means shifts and variance changes for sample size (65, 65), 100,000 repetitions, $\alpha = 0.05$

	Mood-Westenberg Chi-squared										
Means	**Variance change**										
shift	**1.00**	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.041	0.486	0.894	0.985	0.998						
0.01	0.047	0.408	0.866	0.985	0.998						
0.02	0.047	0.406	0.866	0.983	0.997						
0.03	0.047	0.389	0.865	0.983	0.997						
0.04	0.047	0.392	0.868	0.975	0.997						
0.05	0.047	0.404	0.839	0.975	0.997						
0.06	0.047	0.404	0.838	0.975	0.997						
0.07	0.048	0.409	0.839	0.975	0.997	0.999					
0.08	0.046	0.413	0.839	0.976	0.997						
0.09	0.058	0.376	0.836	0.977	0.997						
0.10	0.057	0.375	0.833	0.971	0.996						
0.11	0.057	0.302	0.833	0.971	0.996						
0.12	0.058	0.302	0.831	0.966	0.996						
	Siegel-Tukey *Z*-score										
0.00	0.050	0.727	0.988								
0.01	0.039	0.712	0.984								
0.02	0.039	0.711	0.984								
0.03	0.039	0.698	0.984								
0.04	0.040	0.695	0.983								
0.05	0.040	0.663	0.979	0.999							
0.06	0.040	0.664	0.978	0.999							
0.07	0.040	0.649	0.978	0.999							
0.08	0.038	0.651	0.978	0.999							
0.09	0.024	0.634	0.978	0.999							
0.10	0.024	0.634	0.973	0.999							
0.11	0.025	0.602	0.974	0.999							
0.12	0.024	0.600	0.973	0.999							

Table 18. Power rates for one-tailed directional test for exponential distribution, various means shifts and variance changes for sample size (20, 20), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared											
Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.026	0.094	0.273	0.464	0.617	0.725	0.794	0.845	0.879	0.899	0.916
0.01	0.026	0.092	0.264	0.458	0.609	0.721	0.794	0.844	0.879	0.899	0.916
0.02	0.026	0.085	0.258	0.448	0.606	0.716	0.793	0.843	0.875	0.901	0.914
0.03	0.027	0.081	0.248	0.439	0.602	0.713	0.791	0.841	0.875	0.902	0.915
0.04	0.025	0.077	0.240	0.435	0.596	0.710	0.789	0.840	0.876	0.899	0.916
0.05	0.028	0.072	0.234	0.426	0.592	0.707	0.786	0.842	0.874	0.901	0.917
0.06	0.029	0.069	0.226	0.421	0.584	0.704	0.783	0.839	0.873	0.899	0.916
0.07	0.029	0.066	0.220	0.411	0.578	0.698	0.781	0.835	0.874	0.900	0.918
0.08	0.030	0.063	0.212	0.404	0.569	0.693	0.779	0.835	0.873	0.899	0.915
0.09	0.032	0.059	0.204	0.400	0.565	0.693	0.778	0.835	0.873	0.897	0.917
0.10	0.034	0.055	0.197	0.392	0.562	0.685	0.774	0.831	0.871	0.899	0.915
0.11	0.035	0.053	0.191	0.382	0.555	0.683	0.771	0.830	0.869	0.897	0.914
0.12	0.037	0.051	0.186	0.375	0.550	0.677	0.769	0.828	0.869	0.900	0.915
Siegel-Tukey *Z*-score											
0.00	0.049	0.312	0.601	0.777	0.875	0.929	0.956	0.974	0.983	0.988	0.991
0.01	0.042	0.305	0.591	0.774	0.872	0.925	0.957	0.973	0.983	0.988	0.991
0.02	0.040	0.294	0.581	0.768	0.872	0.927	0.955	0.972	0.982	0.987	0.991
0.03	0.035	0.283	0.573	0.763	0.871	0.924	0.957	0.972	0.981	0.988	0.991
0.04	0.030	0.270	0.568	0.761	0.866	0.924	0.956	0.972	0.982	0.988	0.991
0.05	0.029	0.257	0.559	0.754	0.868	0.923	0.955	0.973	0.982	0.988	0.991
0.06	0.025	0.249	0.549	0.749	0.863	0.922	0.953	0.971	0.981	0.987	0.991
0.07	0.022	0.238	0.542	0.746	0.860	0.921	0.953	0.972	0.983	0.987	0.991
0.08	0.020	0.226	0.531	0.741	0.855	0.921	0.953	0.971	0.981	0.987	0.991
0.09	0.018	0.217	0.520	0.735	0.853	0.919	0.952	0.971	0.980	0.987	0.991
0.10	0.016	0.207	0.512	0.730	0.853	0.915	0.951	0.970	0.982	0.988	0.991
0.11	0.014	0.198	0.504	0.727	0.847	0.914	0.950	0.969	0.981	0.987	0.991
0.12	0.013	0.189	0.494	0.718	0.847	0.913	0.949	0.969	0.981	0.987	0.991

Table 19. Power rates for one-tailed directional test for exponential distribution, various means shifts and variance changes for sample size (30, 30), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.069	0.241	0.553	0.782	0.899	0.951	0.976	0.988	0.994	0.996	0.997
0.01	0.069	0.232	0.543	0.772	0.896	0.951	0.975	0.988	0.993	0.996	0.997
0.02	0.071	0.222	0.532	0.765	0.890	0.949	0.975	0.987	0.993	0.996	0.997
0.03	0.071	0.212	0.518	0.760	0.887	0.947	0.974	0.987	0.993	0.996	0.997
0.04	0.073	0.204	0.506	0.752	0.885	0.944	0.973	0.986	0.993	0.996	0.997
0.05	0.075	0.195	0.496	0.745	0.879	0.944	0.972	0.986	0.992	0.996	0.997
0.06	0.078	0.183	0.484	0.736	0.876	0.941	0.972	0.986	0.992	0.996	0.997
0.07	0.081	0.176	0.475	0.729	0.872	0.938	0.970	0.985	0.992	0.996	0.997
0.08	0.084	0.166	0.459	0.720	0.866	0.938	0.969	0.984	0.992	0.996	0.997
0.09	0.087	0.158	0.451	0.713	0.865	0.935	0.969	0.984	0.991	0.995	0.997
0.10	0.092	0.150	0.440	0.705	0.858	0.932	0.967	0.985	0.991	0.995	0.997
0.11	0.097	0.143	0.428	0.697	0.852	0.933	0.968	0.983	0.991	0.995	0.997
0.12	0.102	0.137	0.417	0.687	0.850	0.929	0.965	0.983	0.991	0.995	0.997

Siegel-Tukey *Z*-score

0.00	0.049	0.428	0.768	0.917	0.970	0.989	0.995	0.998	0.999	0.999	
0.01	0.043	0.415	0.761	0.914	0.970	0.988	0.995	0.998	0.999		
0.02	0.038	0.398	0.755	0.910	0.968	0.988	0.995	0.998	0.999		
0.03	0.032	0.382	0.743	0.909	0.968	0.988	0.995	0.998	0.999		
0.04	0.029	0.369	0.734	0.904	0.966	0.987	0.995	0.998	0.999	0.999	
0.05	0.024	0.356	0.724	0.902	0.964	0.988	0.994	0.998	0.999		
0.06	0.021	0.336	0.720	0.897	0.963	0.986	0.995	0.997	0.999	0.999	
0.07	0.018	0.324	0.710	0.893	0.963	0.986	0.995	0.998	0.999		
0.08	0.016	0.306	0.700	0.891	0.961	0.986	0.995	0.998	0.999		
0.09	0.014	0.291	0.690	0.887	0.961	0.985	0.994	0.998	0.999	0.999	
0.10	0.012	0.275	0.678	0.882	0.958	0.985	0.994	0.998	0.999	0.999	
0.11	0.011	0.262	0.667	0.879	0.956	0.985	0.994	0.997	0.999		
0.12	0.009	0.249	0.658	0.875	0.955	0.984	0.994	0.998	0.999	0.999	

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 233-245.
doi: 10.22237/jmasm/1493597520

ISSN 1538 − 9472

Factor Analysis by Limited Scales: Which Factors to Analyze?

Stan Lipovetsky
GfK North America
Minneapolis, MN

Factor Analysis (FA) and Principal Component Analysis (PCA) are well-known main tools of the multivariate statistics for data analysis, reduction, and visualization. Commonly, the analysis and interpretation of their solutions is performed for each of several main eigenvectors with variances explaining a big part of the total variability in data. The recommendation is to determine if all the main vectors are really needed in the analysis, or some of them should be skipped if they correspond to the absence of the analyzing features. A simple criterion for identifying redundant vectors of loadings is their negative correlation with the vector of mean values of the original variables. Limited Likert scales of measurements are considered, and it is shown variables correlations and variances are connected to the mean values. FA and PCA structures defined by subsets of highly related variables can correspond to the lower levels of Likert scales meaning the absence of the measured features, so these loading vectors could be senseless for interpretation. Numerical examples are discussed on marketing research data.

Keywords: FA, PCA, loadings, eigenvectors, interpretation

Introduction

Factor Analysis (FA), Principal Component Analysis (PCA), and also Singular Value Decomposition (SVD) are well-known main tools of the multivariate statistics for data analysis, reduction, and visualization, widely used already for many dozen years (for instance, Lawley & Maxwell, 1971; Timm, 1975; Harman, 1976; Dillon & Goldstein, 1984) and continuing to be described and developed in numerous works (Bartholomew & Knott, 1999; Skrondal & Rabe-Hesketh, 2004; Lipovetsky & Conklin 2005; Elden, 2007; Härdle & Hlávka, 2007; Motoda & Liu, 2008; Izenman, 2008; Härdle & Simar, 2012; Lipovetsky, 2009, 2012, 2015). The analysis and interpretation of their solutions is usually performed for several first

Stan Lipovetsky, Ph.D., is a Senior Research Director in the Marketing Sciences department. Email them at: stan.lipovetsky@gfk.com.

retained eigenvectors with bigger variances explaining a main part of the total variability in data.

Variables defined in Likert scales often applied in marketing research and other social measurements are considered. It is a limited scale of, for instance, four, five, seven, or ten levels for measuring characteristics of interest. The paper shows that the variables' mean values can influence their variances, correlations, and the loadings of FA or PCA. In some cases the FA and PCA loading structures defined by subsets of highly related variables can correspond to the levels of Likert scales which actually indicate the absence of the measured features, so such loading vectors could be redundant for analysis and interpretation. The paper suggests checking correlations of the main eigenvectors with the vector of means, and when some of these correlations are negative the related factors may be skipped from consideration if they correspond not to presence but to absence of the analyzing features.

Relation of Means, Standard Deviations, and Correlations for Limited Scales

Consider data from a real marketing research project on features and qualities of protein snacks and shakes, where 1034 respondents evaluated thirty-five attributes by four-point Likert scales with levels

$$\begin{cases} 4 \text{ - definitely applies to me} \\ 3 \text{ - applies to me somewhat} \\ 2 \text{ - does not really apply to me} \\ 1 \text{ - does not apply to me at all} \end{cases} \quad (1)$$

Table 1 presents descriptive statistics on these attributes: means and standard deviations (std).

The graph of std versus mean values is presented in Figure 1 and shows that standard deviations are smaller if mean values are closer to the margins 1 and 4 of this Likert scale. Note that there are less observations on the lower levels of the scale because respondents in marketing research mostly answer at the "better" side of scales. It is intuitively clear that it should be so, because there is simply no space for volatility when most of observations gravitate to one or another margin of a limited scale. Quadratic regression of standard deviation by mean values yields the model:

$$\text{std} = -0.52 + 1.40\,\text{mean} - 0.30\,\text{mean}^2 \quad (2)$$

where the coefficient of multiple determination R^2 =0.88 and the *F*-statistic of 4264 are big, so the model is of a very high quality.

Finding Pearson's pair correlations between all the attributes and stacking them into one matrix together with the corresponding mean values we can consider how correlations depend on mean values. To make such a consideration more clear, we can find 5% quantiles of the means and correlations and present them on one graph – see Figure 2. It shows that there evidently are two areas of higher correlations related to bigger and to smaller mean values.

Fourth-degree polynomial regression corresponding to the plot in Figure 2 yields the model:

$$\text{cor} = -73.30 + 114.60\,\text{mean} - 66.35\,\text{mean}^2 + 16.97\,\text{mean}^3 - 1.625\,\text{mean}^4 \quad (3)$$

with coefficient of multiple determination R^2 = 0.58 and *F*-statistic 5.1, so the model is of a good quality as well. The smaller std at the margins of the limited scale presented in Figure 1 are translated onto the bigger correlation (as covariance divided by standard deviations of the correlated variables) in Figure 2.

Table 1. Means and std for 35 attributes measure by 4-point Likert scale

attribute	mean	std	attribute	mean	std
1	2.77	1.06	19	2.94	1.03
2	2.86	1.06	20	2.94	1.00
3	2.54	1.08	21	2.06	1.06
4	2.85	1.03	22	2.84	1.03
5	2.81	1.05	23	2.47	1.12
6	2.92	0.99	24	3.08	0.91
7	2.81	1.04	25	2.99	0.98
8	2.91	1.04	26	2.77	1.03
9	2.39	1.08	27	2.27	1.11
10	2.34	1.07	28	3.19	0.90
11	3.07	0.90	29	2.96	0.98
12	3.00	0.98	30	2.99	0.94
13	3.07	0.92	31	2.84	0.97
14	2.97	0.98	32	3.14	0.91
15	2.60	1.08	33	3.03	0.91
16	2.68	1.09	34	2.45	1.11
17	2.78	1.03	35	2.46	1.11
18	1.91	1.08			

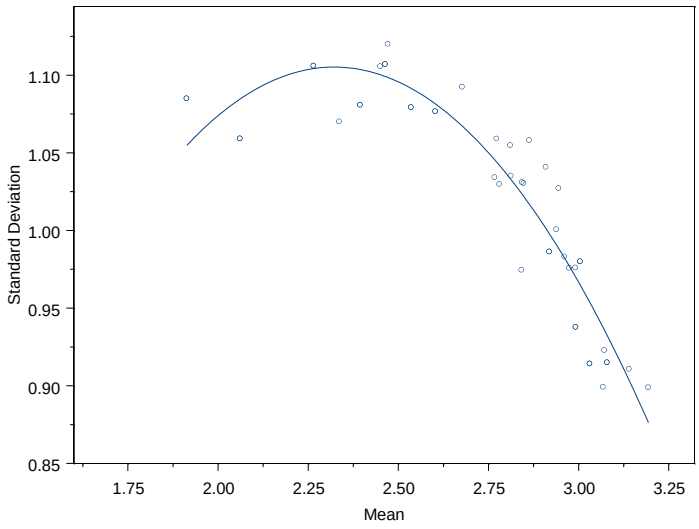


Figure 1. Standard deviation versus mean for attributes measured by Likert 4-point scale

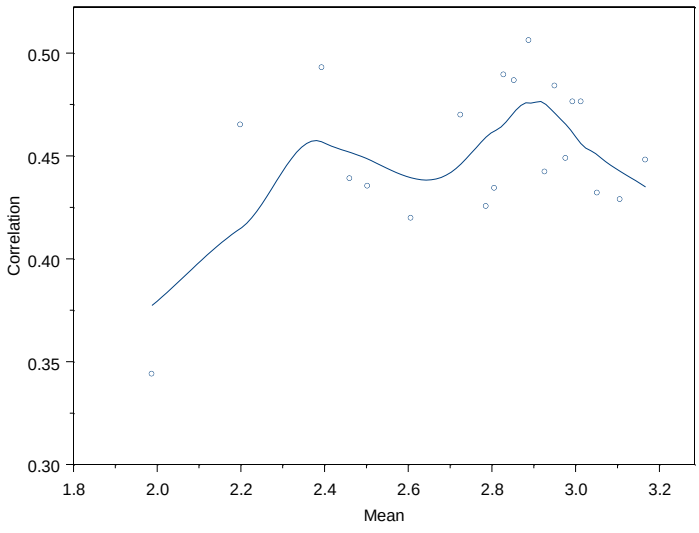


Figure 2. Correlations vs. means for attributes by Likert 4-point scale

Factor Loadings and Their Correlations with Mean Values

Big and approximately equal correlations correspond to the block-diagonal structure of the entire correlation matrix of all variables, where the inter-block correlations are bigger than the outer-block correlations (by absolute value). All pair correlations of the items in this example are positive and varying in the range from 0.35 to 0.55. If some big correlations would be negative, it is always possible to change the variables to the opposite direction by flipping the scale, so all correlations become positive. Let us first briefly describe some results from positive matrix theory.

Due to the Perron-Frobenius theory for a positive matrix's eigenvectors (Salton, 1988; Lipovetsky, 2009; Horn & Johnson, 2013), the first eigenvector of a positive correlation matrix has positive elements and the larger ones identify the variables more related among themselves than with others identified by smaller loadings. Absence of zero elements shows that the matrix is irreducible, or by permutation of variables the matrix cannot be presented in a block-diagonal form when each diagonal block consists of highly correlated subsets of the variables, and the non-diagonal blocks contain zeros. However, higher loadings define a subset of closely-related variables, and the rest of variables with lower loadings could belong to another subset of closely-related variables. In practice, a matrix of correlation can only be approximately presented in a block-diagonal form with higher correlations within the diagonal blocks and with lower correlations in the non-diagonal blocks. If the first eigenvector identifies by the highest elements one of the diagonal blocks, the second eigenvector should correspond to another diagonal block and, due to the Perron-Frobenius theory, it can have positive elements of the variables belonging to this block. The next main eigenvectors can relate to other diagonal blocks and, again, each of them can be flipped by sign.

Let us consider how the results of factor analysis can correspond to different ranges of the mean values shown in Figure 2. FA loadings for 3, 4, and 5-factor solutions obtained in a maximum likelihood approach with additional varimax rotation are presented in Table 2.

The main loadings in Table 2 are colored by dark green. Table 2 also shows the item means, and correlations between them and FA loadings. We see that in each FA solution there is a strong negative correlation of the loadings with mean values of attributes. It can be interpreted as follows.

Table 2. Attribute means, FA loadings, and correlations

		FA-3			FA-4				FA-5				
item	mean	F1	F2	F3	F1	F2	F3	F4	F1	F2	F3	F4	F5
1	2.77	0.12	0.59	0.38	0.60	0.07	0.33	0.22	0.52	0.09	0.33	0.19	0.51
2	2.86	0.13	0.56	0.51	0.55	0.12	0.49	0.18	0.46	0.14	0.51	0.14	0.40
3	2.54	0.20	0.64	0.41	0.64	0.18	0.38	0.19	0.58	0.18	0.43	0.18	0.16
4	2.85	0.32	0.41	0.62	0.41	0.27	0.56	0.31	0.34	0.28	0.59	0.29	0.16
5	2.81	0.24	0.50	0.64	0.49	0.22	0.62	0.24	0.44	0.22	0.65	0.24	0.10
6	2.92	0.73	0.16	0.10	0.14	0.79	0.15	0.04	0.16	0.78	0.15	0.04	-0.03
7	2.81	0.66	0.20	0.01	0.19	0.64	0.00	0.15	0.20	0.64	0.01	0.13	0.05
8	2.91	0.24	0.42	0.68	0.40	0.23	0.69	0.20	0.36	0.22	0.71	0.20	0.06
9	2.39	0.32	0.67	0.34	0.66	0.29	0.32	0.19	0.67	0.26	0.37	0.22	-0.06
10	2.34	0.30	0.67	0.29	0.67	0.26	0.25	0.22	0.65	0.24	0.30	0.23	0.04
11	3.07	0.56	0.21	0.42	0.22	0.44	0.28	0.50	0.18	0.46	0.28	0.48	0.15
12	3.00	0.52	0.27	0.45	0.29	0.35	0.24	0.64	0.27	0.35	0.26	0.64	0.09
13	3.07	0.70	0.11	0.23	0.10	0.65	0.19	0.27	0.10	0.66	0.19	0.25	0.02
14	2.97	0.72	0.14	0.16	0.12	0.77	0.20	0.08	0.12	0.77	0.20	0.06	0.02
15	2.60	0.44	0.52	0.34	0.54	0.34	0.24	0.39	0.53	0.33	0.27	0.40	0.03
16	2.68	0.57	0.18	0.09	0.18	0.56	0.08	0.16	0.20	0.54	0.09	0.17	-0.08
17	2.78	0.46	0.34	0.29	0.35	0.39	0.20	0.34	0.33	0.38	0.22	0.33	0.07
18	1.91	0.09	0.62	0.06	0.63	0.06	0.03	0.10	0.62	0.05	0.07	0.11	0.12
19	2.94	0.24	0.38	0.69	0.35	0.24	0.72	0.18	0.31	0.23	0.74	0.18	0.05
20	2.94	0.26	0.40	0.63	0.39	0.24	0.61	0.23	0.31	0.26	0.63	0.21	0.20
21	2.06	0.35	0.52	0.01	0.52	0.34	0.02	0.06	0.53	0.32	0.06	0.08	-0.02
22	2.84	0.34	0.45	0.60	0.44	0.29	0.56	0.29	0.41	0.28	0.59	0.30	0.00
23	2.47	0.06	0.68	0.27	0.69	0.03	0.23	0.15	0.63	0.04	0.24	0.11	0.45
24	3.08	0.53	0.20	0.48	0.22	0.36	0.28	0.66	0.20	0.36	0.29	0.65	0.06
25	2.99	0.48	0.28	0.52	0.29	0.37	0.40	0.47	0.27	0.37	0.42	0.47	0.00
26	2.77	0.46	0.42	0.46	0.43	0.37	0.36	0.41	0.43	0.35	0.40	0.43	-0.07
27	2.27	0.24	0.69	0.23	0.68	0.25	0.26	0.04	0.69	0.22	0.32	0.06	-0.06
28	3.19	0.40	0.14	0.63	0.14	0.33	0.55	0.39	0.11	0.33	0.55	0.39	0.00
29	2.96	0.67	0.21	0.33	0.20	0.65	0.32	0.23	0.18	0.66	0.32	0.22	0.07
30	2.99	0.63	0.20	0.40	0.20	0.54	0.32	0.39	0.17	0.56	0.32	0.37	0.12
31	2.84	0.61	0.25	0.29	0.26	0.55	0.22	0.31	0.23	0.56	0.23	0.29	0.13
32	3.14	0.59	0.12	0.28	0.12	0.54	0.24	0.27	0.10	0.56	0.23	0.25	0.12
33	3.03	0.55	0.20	0.46	0.22	0.44	0.33	0.48	0.17	0.46	0.33	0.45	0.19
34	2.45	0.21	0.65	0.38	0.64	0.22	0.40	0.09	0.61	0.20	0.45	0.10	0.06
35	2.46	0.25	0.63	0.25	0.63	0.23	0.23	0.17	0.59	0.22	0.27	0.16	0.12
cor		0.56	-0.79	0.49	-0.79	0.48	0.38	0.56	-0.85	0.52	0.31	0.51	0.05

As is well-known, the vectors of loadings in FA, PCA, and SVD, as eigenvectors of eigenproblems for covariance, correlation, or non-centered second-moment matrices, are defined up to an arbitrary normalizing constant – particularly, up to sign change of all their elements that flips the vectors to

opposite direction. It is so for maximum likelihood and other methods of estimation, with orthogonal, oblique, and rotated solutions as well. Negative correlations of some vectors of loading with mean values of attributes can be observed practically in any FA or PCA solution, but it does not eliminate such factors from analysis and interpretation on the basis of this correlation sign only. However, for Likert scales it could indicate that negative correlation of a factor's loadings with the vector of the variables' means occurs because this factor is constituted by the variables with the values mostly on the "lower", or "non-relevant" levels. For instance, such a factor can consist of the attributes getting mostly the lower 1 and 2 levels in the scale of "does not apply to me" meaning in (1).

To check it, let us reshape Table 2 by sorting FA loadings due to the descending order of the items mean values – the results are presented in Table 3. Indeed, it is easy to see by Table 3 that in any FA solution the factors negatively correlated with mean values have the main loadings on the attributes with minimum mean values, in the range below about the mean point 2.5 in the scale (1). But those values correspond to meaningless attributes in this study because they are related to the "non-applied to respondent" levels.

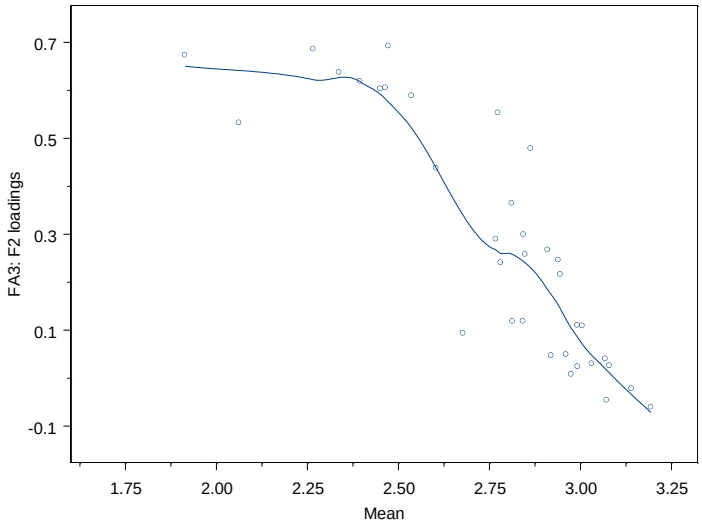


Figure 3. FA-3 solution for 35 attributes with the second factor loadings vs. means

Table 3. Factor loadings sorted by mean values

		FA-3			FA-4				FA-5				
item	mean	F1	F2	F3	F1	F2	F3	F4	F1	F2	F3	F4	F5
28	3.19	0.40	0.14	0.63	0.14	0.33	0.55	0.39	0.11	0.33	0.55	0.39	0.00
32	3.14	0.59	0.12	0.28	0.12	0.54	0.24	0.27	0.10	0.56	0.23	0.25	0.12
24	3.08	0.53	0.20	0.48	0.22	0.36	0.28	0.66	0.20	0.36	0.29	0.65	0.06
11	3.07	0.56	0.21	0.42	0.22	0.44	0.28	0.50	0.18	0.46	0.28	0.48	0.15
13	3.07	0.70	0.11	0.23	0.10	0.65	0.19	0.27	0.10	0.66	0.19	0.25	0.02
33	3.03	0.55	0.20	0.46	0.22	0.44	0.33	0.48	0.17	0.46	0.33	0.45	0.19
12	3.00	0.52	0.27	0.45	0.29	0.35	0.24	0.64	0.27	0.35	0.26	0.64	0.09
25	2.99	0.48	0.28	0.52	0.29	0.37	0.40	0.47	0.27	0.37	0.42	0.47	0.00
30	2.99	0.63	0.20	0.40	0.20	0.54	0.32	0.39	0.17	0.56	0.32	0.37	0.12
14	2.97	0.72	0.14	0.16	0.12	0.77	0.20	0.08	0.12	0.77	0.20	0.06	0.02
29	2.96	0.67	0.21	0.33	0.20	0.65	0.32	0.23	0.18	0.66	0.32	0.22	0.07
19	2.94	0.24	0.38	0.69	0.35	0.24	0.72	0.18	0.31	0.23	0.74	0.18	0.05
20	2.94	0.26	0.40	0.63	0.39	0.24	0.61	0.23	0.31	0.26	0.63	0.21	0.20
6	2.92	0.73	0.16	0.10	0.14	0.79	0.15	0.04	0.16	0.78	0.15	0.04	-0.03
8	2.91	0.24	0.42	0.68	0.40	0.23	0.69	0.20	0.36	0.22	0.71	0.20	0.06
2	2.86	0.13	0.56	0.51	0.55	0.12	0.49	0.18	0.46	0.14	0.51	0.14	0.40
4	2.85	0.32	0.41	0.62	0.41	0.27	0.56	0.31	0.34	0.28	0.59	0.29	0.16
22	2.84	0.34	0.45	0.60	0.44	0.29	0.56	0.29	0.41	0.28	0.59	0.30	0.00
31	2.84	0.61	0.25	0.29	0.26	0.55	0.22	0.31	0.23	0.56	0.23	0.29	0.13
5	2.81	0.24	0.50	0.64	0.49	0.22	0.62	0.24	0.44	0.22	0.65	0.24	0.10
7	2.81	0.66	0.20	0.01	0.19	0.64	0.00	0.15	0.20	0.64	0.01	0.13	0.05
17	2.78	0.46	0.34	0.29	0.35	0.39	0.20	0.34	0.33	0.38	0.22	0.33	0.07
1	2.77	0.12	0.59	0.38	0.60	0.07	0.33	0.22	0.52	0.09	0.33	0.19	0.51
26	2.77	0.46	0.42	0.46	0.43	0.37	0.36	0.41	0.43	0.35	0.40	0.43	-0.07
16	2.68	0.57	0.18	0.09	0.18	0.56	0.08	0.16	0.20	0.54	0.09	0.17	-0.08
15	2.60	0.44	0.52	0.34	0.54	0.34	0.24	0.39	0.53	0.33	0.27	0.40	0.03
3	2.54	0.20	0.64	0.41	0.64	0.18	0.38	0.19	0.58	0.18	0.43	0.18	0.16
23	2.47	0.06	0.68	0.27	0.69	0.03	0.23	0.15	0.63	0.04	0.24	0.11	0.45
35	2.46	0.25	0.63	0.25	0.63	0.23	0.23	0.17	0.59	0.22	0.27	0.16	0.12
34	2.45	0.21	0.65	0.38	0.64	0.22	0.40	0.09	0.61	0.20	0.45	0.10	0.06
9	2.39	0.32	0.67	0.34	0.66	0.29	0.32	0.19	0.67	0.26	0.37	0.22	-0.06
10	2.34	0.30	0.67	0.29	0.67	0.26	0.25	0.22	0.65	0.24	0.30	0.23	0.04
27	2.27	0.24	0.69	0.23	0.68	0.25	0.26	0.04	0.69	0.22	0.32	0.06	-0.06
21	2.06	0.35	0.52	0.01	0.52	0.34	0.02	0.06	0.53	0.32	0.06	0.08	-0.02
18	1.91	0.09	0.62	0.06	0.63	0.06	0.03	0.10	0.62	0.05	0.07	0.11	0.12
cor		0.56	-0.79	0.49	-0.79	0.48	0.38	0.56	-0.85	0.52	0.31	0.51	0.05

So we can see by negative correlations of FA loadings and means that it is possible to identify the variables gravitating to the levels of "does not really apply to me" and "does not apply to me at all". Such attributes do not supply useful information elicited from the respondents. Thus, the factors negatively correlated

with means can be skipped from the analysis and interpretation. For illustration, the loadings of the second factor in the solution with three factors (FA-3 solution, the factor F2 in Table 3) are shown in Figure 3, which clearly describes a negative pattern of correlation.

Cleaning data from inadequate variables always helps to a meaningful statistical analysis, so FA can be re-run without the redundant variables of mostly the irrelevant levels on the limited scale. It is also interesting to note that the PCA loadings even without rotation produce similar to FA correlations with means. For instance, correlations of three first PCA vectors with the vector of means are 0.72, -0.62, and 0.31, so very close to three factor solution's correlations given at the last row in Table 3.

Table 4. Correlations of means and FA loadings for several factor solutions with 45 attributes measure by 7-point Likert scale

	F1	F2	F3	F4	F5	F6
FA-3	0.89	0.08	-0.89			
FA-4	0.87	0.03	-0.87	0.19		
FA-5	0.82	0.14	-0.94	0.21	0.17	
FA-6	0.83	0.15	-0.95	0.21	0.14	0.07

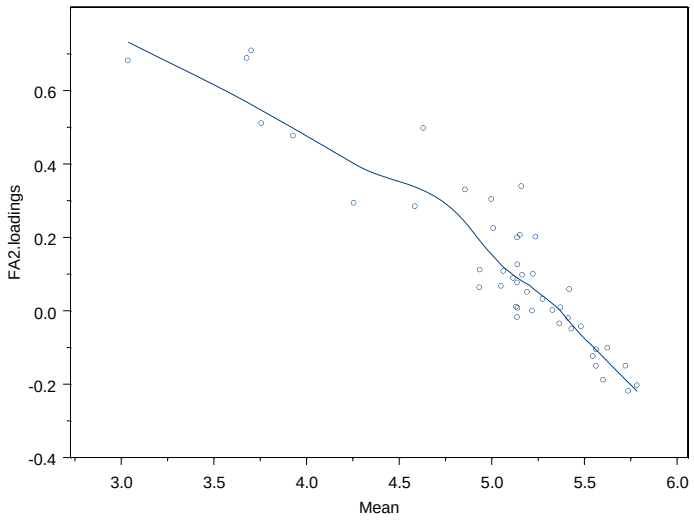


Figure 4. FA-3 solution for 45 attributes with the third factor loadings versus means

Table 5. FA-3 loadings sorted by means for 45 attributes measured by 7-point Likert scale

item	mean	F1	F2	F3	item	mean	F1	F2	F3
21	5.79	0.62	0.19	-0.02	13	5.14	0.48	0.31	0.30
22	5.74	0.75	0.15	-0.01	15	5.14	0.45	0.55	0.28
20	5.72	0.70	0.11	0.04	28	5.14	0.28	0.70	0.20
41	5.63	0.57	0.15	0.07	31	5.14	0.41	0.66	0.20
18	5.60	0.65	0.41	0.04	37	5.14	0.52	0.16	0.35
26	5.57	0.70	0.37	0.13	29	5.13	0.32	0.68	0.21
43	5.56	0.69	0.42	0.09	12	5.12	0.46	0.34	0.26
10	5.55	0.63	0.40	0.10	40	5.07	0.51	0.17	0.26
14	5.48	0.65	0.36	0.18	17	5.05	0.24	0.70	0.25
32	5.43	0.59	0.48	0.18	36	5.01	0.53	0.17	0.38
5	5.42	0.73	0.24	0.28	38	5.00	0.57	0.17	0.47
3	5.41	0.60	0.37	0.19	25	4.94	0.41	0.47	0.29
42	5.37	0.59	0.19	0.19	35	4.93	0.34	0.66	0.26
45	5.37	0.64	0.28	0.17	44	4.86	0.45	0.36	0.50
7	5.33	0.54	0.45	0.21	2	4.63	0.37	0.20	0.62
24	5.28	0.56	0.46	0.25	39	4.59	0.27	0.43	0.42
16	5.24	0.56	0.15	0.36	30	4.26	0.01	0.60	0.40
33	5.23	0.57	0.16	0.27	23	3.93	0.18	0.32	0.57
19	5.22	0.43	0.64	0.22	8	3.76	0.08	0.41	0.60
27	5.19	0.59	0.23	0.23	9	3.70	0.11	0.18	0.76
34	5.17	0.52	0.47	0.31	6	3.68	0.11	0.22	0.74
1	5.16	0.56	0.20	0.51	11	3.04	0.03	0.10	0.70
4	5.15	0.63	0.23	0.40	cor		0.89	0.08	-0.89

In another data set from the same marketing research project, forty five attributes had been measured by a 7-point Likert scale, from 7 meaning "extremely important" to 1 meaning "not at all important." A general structure of the relations between mean values and FA loadings is very similar to that described above for the smaller set of attributes. Table 4 presents the correlations between mean values and factors loadings from three factor solution (FA-3 in the first row) to six factor solution (FA-6 in the last row).

It is useful to note that PCA loading correlated with mean values also yield negative values. PCA constructed by correlation matrix gives three first correlations 0.93, 0.66, and -0.21, and PCA by covariance matrix produces correlations 0.97, -0.29, and -0.11. By Table 4 we see that adding more factors does not change the correlations of the first three factors (F1, F2, and F3 in the first columns) with the mean values of attributes. So, for illustration on the FA loading sorted by means of the attributes, it is sufficient to use the FA-3 solution which is presented in Table 5. This solution demonstrates that the negative

correlation of the loadings with means is observed for the third factor mostly defined by the attributes with mean values below the mid-point of the scale. So there is no need to consider and interpret this 3rd factor defined mostly by the "non-important" attributes. The last factor's loadings for 45 attributes solution from Table 5 is presented in Figure 4 with decreasing loadings profiled by the mean values.

Another interesting example of factor analysis performed on eighty adjectives measured by a 5-point Likert scale for characterizing the beauty of a mathematical proof can be found in Inglis and Aberdein (2014), with the second factor excluded from interpretation because of correspondence to lower levels of description accuracy.

Summary

The work considers the possibility to identify factors which can be skipped from interpretation and further application. The analysis is based on correlations of factor loadings with means of variables constituting the factors. Although the factor and principal component loadings are defined up to their sign, the correlations of factor loadings with variables' means permit the identification of factors consisting mostly of variables measured in Likert scales related to non-relevant values. The variables' means can influence the variances and correlations, which in turn define the factor loadings. In some factors the loading structure defined by subsets of highly-related variables can correspond to the "non-important" levels by Likert scale. Factor loadings after rotation to a simpler structure contain mostly the positive elements, so their negative correlations with the attribute means is a convenient indicator of the redundant factors which can be skipped from further analysis. Thus, depending on the content of a scale levels, there are studies with all main factors making sense, so they can be interpreted and used. Negative correlation of the loadings with mean values of variables in such a case simply shows that lower-level observations define this factor. But on the other hand, there could be studies where factors negatively correlated with mean values can be excluded from consideration because they rather correspond to the absence of analyzing features.

References

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London, UK: Arnold.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York, NY: Wiley.

Elden, L. (2007). *Matrix methods in data mining and pattern recognition*. Philadelphia, PA: SIAM. doi: 10.1137/1.9780898718867

Härdle, W., & Hlávka, Z. (2007). *Multivariate statistics: Exercises and solutions*. New York, NY: Springer. doi: 10.1007/978-3-642-36005-3

Härdle, W. K., & Simar, L. (2012). *Applied multivariate statistical analysis*. New York, NY: Springer. doi: 10.1007/978-3-642-17229-8

Harman, H. H. (1976). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Horn, R. A., & Johnson, C. R. (2013). *Matrix analysis* (2nd ed,). New York, NY: Cambridge University Press.

Inglis, M., & Aberdein, A. (2014). Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica, 23*(1), 87-109. doi: 10.1093/philmat/nku014

Izenman, A. J. (2008). *Modern multivariate statistical techniques*. New York, NY: Springer. doi: 10.1007/978-0-387-78189-1

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. New York, NY: American Elsevier.

Lipovetsky, S. (2009). PCA and SVD with nonnegative loadings. *Pattern Recognition, 42*(1), 68-76. doi: 10.1016/j.patcog.2008.06.025

Lipovetsky, S. (2012). Dual PLS analysis. *International Journal of Information Technology & Decision Making, 11*(05), 879-891. doi: 10.1142/s0219622012500241

Lipovetsky, S. (2015). MANOVA, LDA, and FA criteria in clusters parameter estimation. *Cogent Mathematics, 2*. doi: 10.1080/23311835.2015.1071013

Lipovetsky, S. & Conklin, M. (2005). Singular value decomposition in additive, multiplicative, and logistic forms. *Pattern Recognition, 38*(7), 1099-1110. doi: 10.1016/j.patcog.2005.01.010

Motoda, H., & Liu, H. (Eds.). (2008). *Computational methods of feature selection*. Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9781584888796

Salton, G. (1988). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall/CRC.

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterey, CA: Brooks/Cole.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 246-260.
doi: 10.22237/jmasm/1493597580

ISSN 1538 − 9472

Multivariate Rank Outlyingness and Correlation Effects

Olusola Samuel Makinde
Federal University of Technology
Akure, Nigeria

The effect of correlation on multivariate rank outlyingness, a result of deviation of multivariate rank functions from property of spherical symmetry, is examined. Possible affine invariant versions of this multivariate rank are surveyed, and outlyingness of affine invariant and non-invariant spatial rank functions under general affine transformation are compared.

Keywords: rank function, outlyingness function, symmetry, correlation

Introduction

Ordering of data and the search for the units lying far from the centroid is closely related to searching for outliers in the data cloud. In a univariate setting, this ordering is a linear ranking from smallest to largest. Given sample points $X_1, X_2, \ldots, X_n$, we can order them by their rank values. Ordering of univariate objects based on rank does not depend heavily on the underlying distribution of the data, nor involve estimation of parameters of probability distributions. Similarly in a multivariate setting, we can order multivariate sample points $X_1, X_2, \ldots, X_n$ by their rank function.

An appealing way of working with probability distributions in $\mathbb{R}^d$, especially in nonparametric inference, is through "descriptive measures" that characterize features of particular interest (Serfling, 2004, p. 260). One attractive approach is to base the measures on outlyingness of multivariate rank. In the last couple of decades, notions of multivariate signs and ranks have become a useful tool in analyzing multivariate data, as it does not depend heavily on distributional assumptions, and characterizes the central and extreme observations quite effectively (Makinde & Chakraborty, 2015). Use of multivariate rank for ordering

Olusola Samuel Makinde is a Lecturer in the Department of Statistics. Email at osmakinde@futa.edu.ng.

of data preserves the direction of the data. Möttönen & Oja (1995), Möttönen, Oja & Tienari (1997) used the notion of spatial ranks to construct multivariate tests of location.

A related notion to multivariate ranks is the data depth. Data depth measures depth or centrality of a *d*-dimensional observation with respect to a multivariate data cloud or underlying multivariate distribution. Depth functions in literature include Mahalanobis depth, half-space depth, simplicial depth, likelihood depth, and projection depth, among others. Liu, Parelius & Singh (1999) proposed various ideas on analyzing multivariate data using data depths. We refer readers to Liu, Parelius & Singh (1999) for detailed discussion on depth functions. Statistical approaches based on most of these depth functions suffer computational complexities of the depth functions.

The spatial rank and its outlyingness can be applied in classification and clustering (Makinde, 2015). It has been applied in construction of geometric quantile (Chaudhuri, 1996; Serfling, 2004). It is well known that multivariate rank is not invariant under arbitrary affine transformations, so it may be affected by deviation of population distribution from spherical symmetry. Effect of this deviation on spatial rank outlyingness will be investigated. Based on this, we shall introduce a way of constructing affine invariant multivariate rank outlyingness.

Spatial Rank

Signs and ranks are commonly used in statistical methodology to develop methods or procedures that are independent of distribution assumptions. Use of rank for computing statistical quantities gives robust estimators (e.g. estimator for location) as they are not affected by the presence of outlying values in the data. For the univariate data, *sign* of $x \in \mathbb{R}$ can be defined as

$$sign(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$$

or equivalently,

$$sign(x) = \begin{cases} \dfrac{x}{|x|}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

Univariate centred rank of x with respect to data points $X_1, X_2, \ldots, X_n$ from distribution F can be defined as

$$rank(x) = \frac{1}{n}\sum_{i=1}^{n} sign(x - X_i).$$

Following are some of the basic properties of $rank(x)$,

1. $|\, rank(x)\,| \leq 1$.
2. $|\, rank(x)\,| = 0$ implies x is the median and $|\, rank(x)\,| = 1$ implies x is an extreme point.
3. $E(|\, rank(x)\,|) = 2F(x) - 1$

These properties suggest that $rank(x)$ is not only a useful descriptive statistics, it also characterizes the distribution. Now, we want to define sign and rank functions in a multivariate set up following Chakraborty (2001). Suppose $\boldsymbol{x} \in \mathbb{R}^d$, then the l_p sign of $\boldsymbol{x}$ is

$$sign_p(\boldsymbol{x}) = \begin{cases} \dfrac{\partial}{\partial \boldsymbol{x}} \|\boldsymbol{x}\|_p = \dfrac{\upsilon(\boldsymbol{x})}{\|\boldsymbol{x}\|_p^{\,p-1}}, & \boldsymbol{x} \neq \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{x} = \boldsymbol{0} \end{cases}$$

where

$$\|\boldsymbol{x}\|_p = \left(x_1^p + x_2^p + \cdots + x_d^p\right)^{\frac{1}{p}} \text{ and}$$
$$\upsilon(\boldsymbol{x}) = \left(sign(x_1)|x_1|^{p-1}, \ldots, sign(x_d)|x_d|^{p-1}\right).$$

The l_p rank of $\boldsymbol{x} \in \mathbb{R}^d$ with respect to data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ is defined as

$$rank_p(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} sign_p(\boldsymbol{x} - \boldsymbol{X}_i).$$

when $p = 1$, $sign(\boldsymbol{x}) = \left(sign(x_1), sign(x_2), \ldots, sign(x_d)\right)^T$, the vector of co-ordinatewise signs and for $p = 2$,

$$sign_2(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$$

where $\|.\|_2$ is the Euclidean norm defined as $\|\boldsymbol{y}\|_2 = \left(y_1^2 + y_2^2 + \cdots + y_d^2\right)^{\frac{1}{2}}$. $sign_2(\boldsymbol{x})$ is called the spatial sign vector.

Suppose $\boldsymbol{X}$ is a d-dimensional random vector having a distribution F, which is assumed to be absolutely continuous with respect to the Lebesgue measure $\mathbb{R}^d$. The spatial rank function (Möttönen & Oja, 1995) of any point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to F is defined as

$$rank_F(\boldsymbol{x}) = E_F\left(\frac{\boldsymbol{x} - \boldsymbol{X}}{\|\boldsymbol{x} - \boldsymbol{X}\|}\right). \tag{1}$$

Here $\|.\|$ is the usual Euclidean norm. It follows immediately from the definition that $rank_F(\boldsymbol{x}) = \boldsymbol{0}$ implies that $\boldsymbol{x}$ is the spatial median of the multivariate distribution F. Koltchinskii (1997) established that this spatial rank function is a one-to-one function of the distribution function F and hence it characterizes the distribution. Moreover the direction of the vector $rank_F(\boldsymbol{x})$ suggests the direction in which $\boldsymbol{x}$ is extreme compared to the distribution. Using this idea, Serfling (2004) introduced $\|rank_F(\boldsymbol{x})\|$ as a measure of outlyingness and defined several descriptive measures. Smaller values of $\|rank_F(\boldsymbol{x})\|$ implies that $\boldsymbol{x}$ is more central to the distribution and larger values of $\|rank_F(\boldsymbol{x})\|$ indicates that $\boldsymbol{x}$ is more extreme. If $\|rank_F(\boldsymbol{x})\| = 0$, then $\boldsymbol{x}$ is the spatial median.

Spatial rank helps determine the geometric position of points in $\mathbb{R}^d$ with respect to the data cloud, and hence can be viewed as a descriptive statistic (Guha, 2012). Suppose F is spherically symmetric and characterized by location parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, $\|rank_F(\boldsymbol{x})\|$ increases as $\|\boldsymbol{x} - \boldsymbol{\theta}\|$ increases. This result is stated formally in Theorem 1 below:

Theorem 1. If $\boldsymbol{x}$ has spherically symmetric distribution F with $\boldsymbol{\theta}$ as the centre of symmetry, then for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$rank_F(\boldsymbol{x}) = q\left(\|\boldsymbol{x}-\theta\|\right)\frac{\boldsymbol{x}-\theta}{\|\boldsymbol{x}-\theta\|}$$

for some increasing, non-negative function q.

This is proved in Guha (2012). Following Theorem 1, smaller rank outlyingness indicates more central observation and larger rank outlyingness indicates extreme observation. The following results hold for rank outlyingness:

Fact: Let $\|rank_F(\boldsymbol{x})\|$ denote the measure of outlyingness of $rank_F(\boldsymbol{x})$. Then

1. $\|rank_F(\boldsymbol{x} + \boldsymbol{\theta})\| = \|rank_F(\boldsymbol{x})\|$ for a constant vector $\boldsymbol{\theta}$.
2. $\|rank_F(\boldsymbol{A}\boldsymbol{x})\| = \|rank_F(\boldsymbol{x})\|$ for an orthogonal matrix $\boldsymbol{A}$.

The first expression above implies that rank outlyingness is invariant under location shift or translation while the second indicates that rank outlyingness is invariant under orthogonal scale transformation. In practice, the rank functions $rank_F$ will hardly be known completely and we need to estimate them from the training sample. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be a random sample from a population having distribution F. We define the empirical rank function as

$$rank_{F_n}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x}-\boldsymbol{X}_i}{\|\boldsymbol{x}-\boldsymbol{X}_i\|}$$

Theorem 2. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent and identically distributed d-dimensional random vectors having distribution function F, which is absolutely continuous, then as $n \to \infty$,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\left\|rank_{F_n}(\boldsymbol{x})\right\| - \left\|rank_F(\boldsymbol{x})\right\|\right| \to 0.$$

The proof follows from Koltchinskii's (1997) work on the convergence of the empirical version of spatial rank to its population analogue.

Chaudhuri (1996) defined spatial quantiles as vectors in $\mathbb{R}^d$ that are indexed by a vector $\boldsymbol{u}$ in d-dimensional unit ball. Define an open ball

$B^d = \{\boldsymbol{u} \mid \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\| < 1\}$. For any $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{t} \in \mathbb{R}^d$, also define $\Phi(\boldsymbol{u}, \boldsymbol{t}) = \|\boldsymbol{t}\| + \langle \boldsymbol{u}, \boldsymbol{t} \rangle$, where $\langle .,. \rangle$ denotes the usual Euclidean inner product. Spatial quantile corresponding to $\boldsymbol{u}$ and based on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ is defined as

$$\hat{Q}_n(\boldsymbol{u}) = \arg\min_{Q \in \mathbb{R}^d} \sum_{i=1}^{n} \Phi(\boldsymbol{u}, \boldsymbol{X}_i - Q).$$

It follows from Theorem 1.1.2 of Chaudhuri (1996) that

$$\sum_{i=1}^{n} \frac{Q_n(\boldsymbol{u}) - \boldsymbol{X}_i}{\|Q_n(\boldsymbol{u}) - \boldsymbol{X}_i\|} + n\boldsymbol{u} = 0$$

if $Q_n(\boldsymbol{u}) \neq \boldsymbol{X}_i$ for all $1 \leq i \leq n$. This implies

$$\boldsymbol{u} = \frac{1}{n} \sum_{i=1}^{n} \frac{Q_n(\boldsymbol{u}) - \boldsymbol{X}_i}{\|Q_n(\boldsymbol{u}) - \boldsymbol{X}_i\|}. \tag{2}$$

Serfling (2004) defined $rank_{F_n}(\boldsymbol{x})$ as the inverse function of the spatial quantile function, $\hat{Q}_n(\boldsymbol{u})$. Mathematically, we can write (2) as $\boldsymbol{u} = rank_{F_n}(\hat{Q}_n(\boldsymbol{u})) = rank_{F_n}(\boldsymbol{x})$ and so $\hat{Q}_n(\boldsymbol{u}) = \boldsymbol{x}$ implies $rank_{F_n}(\boldsymbol{x}) = \boldsymbol{u}$. It follows that $rank_{F_n}(\boldsymbol{x})$ is the inverse function of the multivariate geometric quantile function $Q_n(\boldsymbol{u})$ in the sense that $rank_{F_n}(\boldsymbol{x}) = \boldsymbol{u}$ implies that $Q_n(\boldsymbol{u}) = \boldsymbol{x}$ and vice-versa.

Effect of correlation on rank outlyingness

The distribution of a random variable $\boldsymbol{X}$ is said to be spherically symmetric about a parameter $\boldsymbol{\theta}$ if, for any orthogonal matrix $\boldsymbol{B}$,

$$\boldsymbol{X} - \boldsymbol{\theta} \overset{d}{=} \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{\theta})$$

The density function of any spherically symmetric distribution of a random variable $\boldsymbol{X}$, if it exists, is of the form $f(\boldsymbol{x}) \propto g\left((\boldsymbol{x} - \boldsymbol{\theta})^T (\boldsymbol{x} - \boldsymbol{\theta})\right)$ for some

nonnegative scalar function $g(.)$. Similarly, the distribution of a random vector $\boldsymbol{X}$ is said to be elliptically symmetric about $\boldsymbol{\theta}$ if there exists a $d \times d$ nonsingular matrix $\boldsymbol{A}$ such that $\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\theta})$ has a spherically symmetric distribution about $\mathbf{0}$. See Liu (1990), Liu & Singh (1993), Liu, Parelius & Singh (1999) and Serfling (2006) for further reading on multivariate symmetry. The deviation of rank outlyingness from the property of spherical symmetry implies that there exists correlation among variables in the population from which the sample is drawn.

Now, examine the effect of correlation among variables on rank outlyingness. Define $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ and $\boldsymbol{Y}_i = \boldsymbol{A}\boldsymbol{X}_i + \boldsymbol{b}$ for nonsingular matrix $\boldsymbol{A}$ and constant vector $\boldsymbol{b}$, then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{y}-\boldsymbol{Y}_i}{\left\|\boldsymbol{y}-\boldsymbol{Y}_i\right\|} = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{X}_i\right)}{\left\|\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{X}_i\right)\right\|} \neq \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\boldsymbol{x}-\boldsymbol{X}_i\right)}{\left\|\boldsymbol{x}-\boldsymbol{X}_i\right\|}. \tag{3}$$

Table 1. Descriptive statistics of rank outlyingness of bivariate normal objects, bivariate Laplace objects and bivariate t objects with 3 degrees of freedom.

		$\delta = 0$				$\delta = 2$			
	Statistics	$\rho = 0$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.9$
Bivariate normal distribution	*Minimum*	0.0378	0.0272	0.0156	0.0048	0.0799	0.0806	0.0809	0.0804
	25% quantile	0.4396	0.4136	0.4143	0.3739	0.4497	0.4430	0.4258	0.3870
	Median	0.6263	0.6405	0.6157	0.5794	0.6069	0.5900	0.5774	0.5503
	Mean	0.6021	0.5986	0.5873	0.5693	0.6053	0.6007	0.5900	0.5711
	75% quantile	0.7827	0.7852	0.7767	0.7665	0.7948	0.7724	0.7408	0.7524
	Maximum	0.9647	0.9649	0.9846	0.9941	0.9637	0.9678	0.9714	0.9900
Bivariate Laplace distribution	*Minimum*	0.0687	0.0673	0.0589	0.0607	0.0459	0.0588	0.0655	0.0732
	25% quantile	0.4346	0.4429	0.4114	0.3797	0.3688	0.3693	0.3749	0.3770
	Median	0.6133	0.6076	0.5717	0.5410	0.6244	0.6089	0.5749	0.5691
	Mean	0.5952	0.5894	0.5791	0.5649	0.5934	0.5868	0.5762	0.5618
	75% quantile	0.7611	0.7646	0.7821	0.7742	0.7986	0.7942	0.7853	0.7664
	Maximum	0.9693	0.9763	0.9800	0.9832	0.9819	0.9925	0.9955	0.9976
Bivariate *t* distribution with 3 d.f.	*Minimum*	0.1054	0.1129	0.1050	0.0871	0.0883	0.0865	0.0899	0.0698
	25% quantile	0.4076	0.4075	0.3900	0.3569	0.4260	0.4158	0.4098	0.3951
	Median	0.6188	0.5967	0.5705	0.5433	0.6054	0.6009	0.5817	0.5566
	Mean	0.5940	0.5849	0.5716	0.5546	0.5945	0.5885	0.5783	0.5630
	75% quantile	0.8034	0.7875	0.7682	0.7656	0.7734	0.7715	0.7890	0.7600
	Maximum	0.9833	0.9843	0.9927	0.9978	0.9948	0.9964	0.9986	0.9996

As illustration of the effect of correlation on rank outlyingness in (3), a small simulation study is presented. Consider a population to be bivariate elliptically symmetric with centre of symmetry $\boldsymbol{\mu} = (\delta \ 0)^T$ and scale matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Simulate a random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, where sample size n is taken to be 100, and estimate the rank outlyingness function. For various values of ρ, Table 1 presents rank outlyingness for bivariate normally distributed sample, bivariate Laplace distributed sample and bivariate t distributed sample with 3 degrees of freedom.

The outlyingness function behaves anomalously for different values of $\rho \in [0,1)$ irrespective of class distribution. For each family of distribution, descriptive statistics are not in any specific order of ρ. The reason is that though the distribution of $\boldsymbol{X}_i$ is taking more ellipsoid form as ρ increases, the rank outlyingness is being computed with respect to sphere as spatial rank is non-invariant under affine transformation. To overcome the problem of affine non-invariance property of spatial rank, affine invariant versions of rank outlyingness are suggested next.

Affine Invariant Rank Function

Approach based on Cholesky decomposition of the covariance matrix

Spatial rank function can also be defined (Makinde & Chakraborty, 2015) as

$$rank_F^*(\boldsymbol{x}) = E_F\left(\frac{\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{X})}{\left\|\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{X})\right\|}\right)$$

where $\boldsymbol{V}$ is a $d \times d$ matrix such that $\boldsymbol{VV^T} = \boldsymbol{c\Sigma}$ for some constant $\boldsymbol{c}$. If the covariance of the distribution F exists, we can take $\boldsymbol{V}$ to be the Cholesky decomposition of the covariance matrix. For the empirical versions, one can estimate $\boldsymbol{\Sigma}$ by minimum covariance determinant (MCD) estimator of Rousseeuw (1984) and then $\boldsymbol{V}$ by its square root matrix. Note that, the Choleski decomposition of $\boldsymbol{\Sigma}$ (or, its estimate) may not produce an affine invariant rank function but the outlyingness function $\left\|rank_F^*(\boldsymbol{x})\right\|$ will be affine invariant (Makinde & Chakraborty, 2015).

Transformation and re-transformation approach

Chakraborty & Chaudhuri (1996) proposed transformation and re-transformation methodology for conversion of non-equivariant and non-invariant measures under affine transformation to affine equivariant and affine invariant versions respectively, using data driven coordinate system. and then used to construct an affine equivariant median. This technique was also used in Chakraborty & Chaudhuri (1998) to construct robust estimate of location; in Chakraborty, Chaudhuri & Oja (1998) to construct an affine equivariant median and angle test; in Chakraborty (2001) to construct an affine equivariant quantile and also in Dutta & Ghosh (2012); and in Makinde & Chakraborty (2015) to construct affine invariant classifier. The concept is to form an appropriate data driven coordinate system and express all the data points in terms of the new coordinate system. Then compute the spatial rank of the transformed data. Define

$$S_n = \left\{ \alpha \mid \alpha \subset \{1, 2, \ldots, n\} \text{ and } |\alpha| = d + 1 \right\}$$

as the collection of all $d + 1$ subset of $\{1, 2, ..., n\}$. For a fixed $\alpha = \{i_0, i_1, \ldots, i_d\} \subset S_n$, we define $\boldsymbol{X}(\alpha)$ to be a $d \times d$ matrix whose columns are $X_{i_1} - X_{i_0}, X_{i_2} - X_{i_0}, \ldots, X_{i_d} - X_{i_0}$. That is, one of the $d + 1$ data points determines the origin and the lines joining that origin to the remaining d data point will form the coordinate system.

Assuming that elements of α are naturally ordered and that $\boldsymbol{X}_i$'s are independent and identically distributed observations with common probability distribution, which is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$, $\boldsymbol{X}(\boldsymbol{\alpha})$ is invertible with probability one (Chakraborty, 2001). So, $\boldsymbol{X}(\boldsymbol{\alpha})$ is the transformation matrix and for each $i \notin \alpha$, the data set $\boldsymbol{X}_i$ is transformed into a new coordinate system, $\boldsymbol{Y}_i = \{\boldsymbol{X}(\alpha)\}^{-1}\boldsymbol{X}_i$ and then compute the rank of $\boldsymbol{y} = \{\boldsymbol{X}(\alpha)\}^{-1}\boldsymbol{x}$. $\boldsymbol{X}(\alpha)$ is chosen in such a way that the columns of $\Sigma^{-\frac{1}{2}}\boldsymbol{X}(\alpha)$ are as orthogonal as possible. Because population covariance matrix $\boldsymbol{\Sigma}$ is unknown in practice, compute its estimate from the data. The choice of α depends on the value of α that minimizes

$$\zeta(\alpha) = \frac{trace\left(\left\{\boldsymbol{X}(\alpha)\right\}^T \Sigma^{-1} \boldsymbol{X}(\alpha)\right)/d}{\det\left(\left\{\boldsymbol{X}(\alpha)\right\}^T \Sigma^{-1} \boldsymbol{X}(\alpha)\right)^{1/d}}$$

so that $\zeta(\alpha)$ becomes very close to 1. Obviously, once α is selected, the computation of affine invariant spatial rank is straightforward in any dimension.

The affine invariant spatial rank is defined as

$$rank_F(\boldsymbol{x}) = E\left(\frac{\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}]}{\left\|\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}]\right\|}\right) \tag{4}$$

The sample version is defined as

$$rank_{F_n}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}_i]}{\left\|\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}_i]\right\|} \tag{5}$$

Suppose $\boldsymbol{X}_i$, $1 \le i \le n$ be samples on $\mathbb{R}^d$ from a distribution F, it is easy to show that the rank function (defined in (5) above) of a data point $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is $rank_G(\boldsymbol{y}) = rank_F(\boldsymbol{x})$, where G is the distribution of $\boldsymbol{y}$. This is shown by the theorem below.

Theorem 3. Suppose $\boldsymbol{X}_i$, $1 \le i \le n$ is a sample on $\mathbb{R}^d$ having a distribution F. For any $\alpha \in S_n$, $rank_{F_n}(\boldsymbol{x})$ defined in (5) is affine invariant.

Hence, the transformed multivariate rank is invariant under affine transformation. Any statistic based on this transformed rank is affine invariant and can handle the problem associated with deviation from spherical symmetry. Gao (2003**)** defined another version of spatial depth based on rank outlyingness defined in (1) and can be made affine invariant by replacing outlyingness of the rank function in (1) by its affine invariant version.

Numerical Example

To illustrate these methodologies, an example based on ordering of iris data (Fisher, 1936) is presented and quantiles of outlyingness functions of the variants of multivariate rank for the three species of iris flower are compared. The species are iris setosa, iris versicolor and iris virginica.

Presented in Table 2 are the quantiles and mean of the outlyingness of affine invariant and non-affine invariant rank for three species of iris data. The data is available on package R. We denote outlyingness function of affine invariant multivariate rank based on Cholesky decomposition of the covariance matrix by CD approach, outlyingness function of affine invariant multivariate rank based on transformation and re-transformation approach by TR approach and outlyingness function of affine non-invariant multivariate rank defined in equation (1) by non-invariant.

Observe that quantiles of rank outlyingness based on Cholesky decomposition of the covariance matrix and one based on transformation and re-transformation approach are close to each but far away from corresponding quantiles of values of outlyingness based on affine non-invariant multivariate rank. The implication of this is that correlation among the four variables (sepal length, sepal width, petal length and petal width) of each observation in the data can affect the performance of any statistical method or test based on non-affine invariant rank outlyingness.

Table 2. Ordering of species of Iris data based on the outlyingness functions of affine invariant and non-affine invariant ranks.

Iris Species	Approaches	*Minimum*	*1st Quartile*	*Median*	*Mean*	*3rd Quartile*	*Maximum*
Setosa	*CD approach*	0.2461	0.5500	0.6805	0.6447	0.7740	0.8827
	TR approach	0.2456	0.5383	0.6792	0.6437	0.7791	0.8820
	Non-invariant	0.1398	0.5138	0.6372	0.6160	0.7640	0.9436
Versicolor	*CD approach*	0.2727	0.5520	0.6811	0.6506	0.7682	0.8755
	TR approach	0.2710	0.5532	0.6862	0.6485	0.7603	0.8676
	Non-invariant	0.2992	0.4759	0.6406	0.6197	0.7317	0.9116
Virginica	*CD approach*	0.3879	0.5578	0.6724	0.6543	0.7382	0.8877
	TR approach	0.3513	0.5394	0.6722	0.6531	0.7596	0.9063
	Non-invariant	0.2808	0.4850	0.6447	0.6204	0.7439	0.9538

Observe that range of outlyingness of observations is noticeably bigger in affine non-invariant rank compare to the affine invariant rank. The minimum outlyingness value is least in affine non-invariant rank and may therefore mis-identify an observation as outlying. Hence, both affine invariant rank outlyingness functions perform quite well.

Conclusion

The effect of correlation on spatial rank outlyingness was considered and its possible applications. The spatial rank outlyingness based on the training sample does not depend on any distributional assumption and does not require any estimation of model parameters. These give a nonparametric flavor to any statistical technique based on multivariate rank. It is also computationally simple and can be applied to very high dimensional data as well. The rank outlyingness is not affine invariant and as a remedial measure we suggested a transformation of the data to a new coordinate system to make the rank outlyingness affine invariant.

The first idea of transformation is based on transformation retransformation approach proposed by Chakraborty (2001). This makes the spatial ranks affine invariant and hence the rank outlyingness becomes affine invariant. The other transformation considered is based on the square root of the scale matrix $\mathbf{\Sigma}$. It requires the estimation of $\mathbf{\Sigma}$ and may result in a non-robust rank outlyingness. Though the resulting spatial ranks are not affine invariant, rank outlyingness is affine invariant and usually computationally very simple if we use the sample covariance matrix as an estimate of $\mathbf{\Sigma}$. When variables of the data are independent of one another, then both affine invariant versions of rank outlyingness reduces to the usual rank outlyingness.

References

Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics, 53*(2), 380–403. doi: 10.1023/a:1012478908041

Chakraborty, B. & Chaudhuri, P. (1996). On a transformation and retransformation technique for constructing affine equivariant multivariate median. *Proceedings of the American Mathematical Society, 124*(8), 2539–2546. doi: 10.1090/s0002-9939-96-03657-x

Chakraborty, B. & Chaudhuri, P. (1998). On an adaptive transformation and retransformation estimate of multivariate location. *Journal of the Royal Statistical Society: Series B, 60*(1), 145–157. doi: 10.1111/1467-9868.00114

Chakraborty B., Chaudhuri, P. & Oja, H. (1998). Operating transformation and re-transformation on spatial median and angle test. *Statistica Sinica, 8*, 767–784.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of American Statistical Association, 91*(434), 862–872. doi: 10.1080/01621459.1996.10476954

Dutta, S. & Ghosh, A. K. (2012). *On classification based on L_p depth with an adaptive choice of p*. Technical Report No. R5/2011, Statistics and Mathematics Unit. Indian Statistical Institute, Kolkata, India

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Gao, Y. (2003). Data depth based on spatial rank. *Statistics & Probability Letters, 65*(3), 217 – 225. doi: 10.1016/j.spl.2003.06.003

Guha, P. (2012). *On scale-scale curves for multivariate data based on rank regions*. PhD thesis, University of Birmingham, UK.

Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. *The Annals of Statistics, 25*(2), 435 – 477. doi: 10.1214/aos/1031833659

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics, 18*(1), 405–414. doi: 10.1214/aos/1176347507

Liu, R. Y. & Singh, K. (1993). A quality index based on multivariate data depth and multivariate rank tests. *Journal of the American Statistical Association, 88*(421), 252–260. doi: 10.1080/01621459.1993.10594317

Liu, R. Y., Parelius, J. M. & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics, 27*(3), 783–858. doi: 10.1214/aos/1018031260

Makinde, O. S. (2015). *On some classification methods for high dimensional and functional data*. PhD Thesis, University of Birmingham

Makinde, O. S. & Chakraborty, B. (2015). On some classifiers based on distribution functions of multivariate ranks. In Nordhausen, K and Taskinen, S. (Eds). *Modern Nonparametric, Robust and Multivariate Methods, Festschrift in Honour of Hannu Oja*. NY: Springer, 249–264. doi: 10.1007/978-3-319-22404-6_15

Möttönen, J. & Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics, 5*, 201–213.

Möttönen, J., Oja, H. & Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *The Annals of Statistics, 25*(2), 542–552. doi: 10.1214/aos/1031833663

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880. doi: 10.1080/01621459.1984.10477105

Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference, 123*(2), 259–278. doi: 10.1016/s0378-3758(03)00156-3

Serfling, R. (2006). Multivariate symmetry and asymmetry, In S. Kotz, N. Balakrishnan, C. B. Read & B. Vidakovic, Eds. *Encyclopedia of Statistical Sciences, Second Ed., 8,* 5338–5345. NY: Wiley. doi: 10.1002/0471667196.ess5011.pub2

Appendix

Proof of Theorem 3. For any $d \times d$ nonsingular matrix $\boldsymbol{A}$, let $\boldsymbol{Y}_i = \boldsymbol{A}\boldsymbol{X}_i + \boldsymbol{b}$. Since $\boldsymbol{X}(\alpha)=[\boldsymbol{X}_{i1} - \boldsymbol{X}_{i0}, \boldsymbol{X}_{i2} - \boldsymbol{X}_{i0}, ..., \boldsymbol{X}_{id} - \boldsymbol{X}_{i0}]$, we have

$$\begin{aligned}
\boldsymbol{Y}(\alpha) &= \left[\boldsymbol{Y}_{i1} - \boldsymbol{Y}_{i0}, \boldsymbol{Y}_{i2} - \boldsymbol{Y}_{i0}, \ldots, \boldsymbol{Y}_{id} - \boldsymbol{Y}_{i0}\right] \\
&= \left[\boldsymbol{A}\boldsymbol{X}_{i1} + \boldsymbol{b} - \left(\boldsymbol{A}\boldsymbol{X}_{i0} + \boldsymbol{b}\right), \boldsymbol{A}\boldsymbol{X}_{i2} + \boldsymbol{b} - \left(\boldsymbol{A}\boldsymbol{X}_{i0} + \boldsymbol{b}\right), \ldots, \boldsymbol{A}\boldsymbol{X}_{id} + \boldsymbol{b} - \left(\boldsymbol{A}\boldsymbol{X}_{i0} + \boldsymbol{b}\right)\right] \\
&= \left[\boldsymbol{A}\boldsymbol{X}_{i1} - \boldsymbol{A}\boldsymbol{X}_{i0}, \boldsymbol{A}\boldsymbol{X}_{i2} - \boldsymbol{A}\boldsymbol{X}_{i0}, \ldots, \boldsymbol{A}\boldsymbol{X}_{id} - \boldsymbol{A}\boldsymbol{X}_{i0}\right] \\
&= \boldsymbol{A}\left[\boldsymbol{X}_{i1} - \boldsymbol{X}_{i0}, \boldsymbol{X}_{i2} - \boldsymbol{X}_{i0}, \ldots, \boldsymbol{X}_{id} - \boldsymbol{X}_{i0}\right] \\
&= \boldsymbol{A}\boldsymbol{X}(\alpha)
\end{aligned}$$

The transformed multivariate rank of a data point $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, where $\boldsymbol{x} \in \mathbb{R}^d$ is

$$
\begin{aligned}
rank_{G_n}\left(\boldsymbol{y}\right) &= rank_{G_n}\left(\boldsymbol{AX}+\boldsymbol{b}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\left\{\boldsymbol{Y}\left(\alpha\right)\right\}^{-1}\left[\boldsymbol{y}-\boldsymbol{Y}_i\right]}{\left\|\left\{\boldsymbol{Y}\left(\alpha\right)\right\}^{-1}\left[\boldsymbol{y}-\boldsymbol{Y}_i\right]\right\|} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\left\{\boldsymbol{AX}\left(\alpha\right)\right\}^{-1}\left[\boldsymbol{Ax}+\boldsymbol{b}-\left[\boldsymbol{AX}_i+\boldsymbol{b}\right]\right]}{\left\|\left\{\boldsymbol{AX}\left(\alpha\right)\right\}^{-1}\left[\boldsymbol{Ax}+\boldsymbol{b}-\left[\boldsymbol{AX}_i+\boldsymbol{b}\right]\right]\right\|} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\left\{\boldsymbol{X}\left(\alpha\right)\right\}^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}\left[\boldsymbol{x}-\boldsymbol{X}_i\right]}{\left\{\boldsymbol{X}\left(\alpha\right)\right\}^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}\left[\boldsymbol{x}-\boldsymbol{X}_i\right]} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\left\{\boldsymbol{X}\left(\alpha\right)\right\}^{-1}\left[\boldsymbol{x}-\boldsymbol{X}_i\right]}{\left\{\boldsymbol{X}\left(\alpha\right)\right\}^{-1}\left[\boldsymbol{x}-\boldsymbol{X}_i\right]} = rank_{F_n}\left(\boldsymbol{x}\right).
\end{aligned}
$$

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 261-278.
doi: 10.22237/jmasm/1493597640

ISSN 1538 − 9472

Outlier Impact and Accommodation on Power

Hongjing Liao
Beijing Foreign Studies University
Beijing, China

Yanju Li
Western Carolina University
Cullowhee, NC

Gordon P. Brooks
Ohio University
Athens, OH

The outliers' influence on power rates in ANOVA and Welch tests at various conditions was examined and compared with the effectiveness of nonparametric methods and Winsorizing in minimizing the impact of outliers. Results showed that, considering both power and Type I error, a nonparametric test is the safest choice to control the inflation of Type I error with a decent sample size and yield relatively high power.

Keywords: Outlier, Monte Carlo simulation, nonparametric, Winsorizing, Type I error, power

Introduction

Outliers are defined as "observations (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (Barnett & Lewis, 1994, p. 4). They are often present in datasets of educational research, and could have disproportionate influence on statistical conclusions. Therefore, outlier detection and outlier treatment have become important issues in the practice of statistical analysis (Bakker, & Wicherts, 2014; Rousseeuw & van Zomeren, 1990). Detection of outliers has been the focus of outlier research for decades, and there is abundant literature on outlier detecting approaches (Berkane & Bentler, 1988; Barnett & Lewis, 1994; Cook, 1986; Gnanadesikan, 1997; Jarrell, 1991).

In practice, the most widely used method is to detect an outlier using the absolute *Z* value in standard normal distributions; a threshold value of *Z* beyond 3 is often used. Other methods include using the median absolute deviation statistic (MAD), the interquartile range (IQR), and different kinds of residuals (Bakker, & Wicherts, 2014; Barnett & Lewis, 1994; Berkane & Bentler, 1988; Cook, 1986; Gnanadesikan, 1997; Jarrell, 1991). There are also bivariate and multivariate techniques for outlier detection, such as principal components, hat matrix, minimum volume ellipsoid, minimum covariance determinant, minimum

Dr. Hongjing Liao is a lecturer. Email her at: hongjing.liao@139.com.

generalized variance, and Mahalanobis distance (Hawkins, 1974; Hoaglin & Welsch, 1978; Stevens, 1984; Wilcox, 2012). Methods of outlier detection may vary depending on research design, methods, and contexts. Yet after detecting an outlier, the researcher faces another challenge of dealing with the outliers. It is suggested that before any treatment on outliers, the unusual observations should be examined and evaluated under the specific context and try to find the reason for their occurrence. Outlier occurrence is usually from the following four sources: (a) errors, such as erroneous data entries, analysis errors, or equipment problems; (b) failure to specify missing values; (c) including a case that does not belong to the target population; (d) an actual value of the target population but the population has more extreme scores than a normal distribution (Freedman, Pisani, & Purves, 2007; Tabachnick & Fidell, 2001; Hampel, 2001; Warner, 2012).

Warner (2012) suggested three approaches of dealing with an outlier: "to retain, omit, or modify" (p. 287). When reasons for outlier occurrence are deterministic, that is, due to apparent errors in execution of data that are controllable, the approach to deal with the outliers is to correct or delete erroneous values. However, when reasons for outlier occurrence are less apparent, it is often recommended to decide on outlier handling before seeing the results of the main analyses and to report transparently about how outliers were handled (Bakker, & Wicherts, 2014; Liao, Li, & Brooks, 2016). Under these circumstances, thoughtless removal of the outliers is often not recommended, as outlying data can be legitimate data points (Orr, Sackett, & DuBois, 1991). When outliers are unusual but substantively meaningful aspects of the intended study, deleting the outliers causes loss of useful information and often increases the probability of finding a false positive (Chow, Hamaker, & Allaire, 2009; Hampel, 2001). If outliers have to be removed, it is suggested to compare the resulting analyses with and without outliers, and then report an assessment of the influence of outliers through deletion (Allison, Gorman, & Primavera, 1993; Bakker, & Wicherts, 2014).

Many other studies suggest outlier accommodation is a more reliable method to address outliers than simple removal (Analytical Methods Committee, 1989). Accommodation of outliers includes using a robust approach to reduce the impact of the outlying observations and treating outliers to lower their impact in statistical tests. Nonparametric statistical ranking is a commonly-used robust test that is shown to be less influenced by outliers; other robust tests also include the Mann-Whitney-Wilcoxon test and the Yuen-Welch test (Zimmerman & Zumbo, 1990).

Other popular approaches to treating outliers include trimming and Winsorizing (Wilcox, 1998; Dixon & Yuen, 1974). Trimming involves removing the extreme values and often results in a loss in sample size and power. Winsorizing is another popular method to reduce the weights of outliers by replacing them with a specific percentile of data-dependent values (Orr, Sackett, & DuBois, 1991). One-end Winsorizing means that, when outliers are all positive or negative, they are replaced from only one end; two-end Winsorizing means replacing outliers from each end. These different approaches of outlier accommodation may well vary in usefulness of producing consistent study results, and may affect both Type I error and power.

Some researchers studied the robustness of nonparametric tests in the presence of outliers (Zimmerman, 1994, 1995; Li et al., 2009), and Zimmerman (1995) found that nonparametric methods based on ranks have an advantage for outlier-prone densities over ANOVA. However, few studies have focused on multiple comparisons of different outlier accommodation methods. In 2014, the authors conducted a Monte Carlo simulation study and examined the influence of outliers on Type I error rates in ANOVA and Welch tests, and compared nonparametric test and Winsorizing at different locations in controlling outlier impact (Liao et al., 2016). In the current study, the authors followed up their previous simulation study to add new approaches to outlier accommodation methods on Type I error, and further explored outlier impact and accommodation methods on power.

Purpose of the Study

The purpose of this study is to look for answers to two practical questions by means of Monte Carlo methods: (1) what is the impact of outliers on statistical power with different effect sizes, sample sizes, and number of outliers? (2) Among the commonly-used outlier accommodation methods, such as nonparametric rank-based test and Winsorizing (one-end and two-end), which method is more effective in reducing outlier impact, and under what circumstances?

In this study, outliers' influence on statistical power in ANOVA and Welch tests were examined with different effect sizes, sample sizes, and number of outliers. Furthermore, two basic approaches in handling outliers, nonparametric tests and Winsorizing, and their effectiveness in controlling outlier impact were investigated. More specifically, the study compared the statistical power in the following two conditions: when the outliers were retained and non-parametric

methods were then applied to the data, and when outliers were treated using Winsorizing. As there has been no consensus regarding the percentile of Winsorizing and little information provided on how to decide the locations in existing literature, this study explored both one-end and two-end Winsorizing, and compared their difference in statistical power and Type I error.

Compared with outlier detection, there are few studies that concentrate on outlier treatment methods and even fewer on comparisons of outlier accommodation techniques. This study ventures to explore some new areas based on existing studies. From the research design perspective, when the reason for outlier occurrence cannot be traced – which frequently happens in statistical analyses of educational research – it is reasonable to retain the outliers but give less weight to their influence. Therefore, understanding the impact brought by the presence of outliers and choosing an appropriate method for outlier accommodation are critical for credible analysis and conclusion. Moreover, this study focused on multiple comparisons of outlier accommodation techniques and presents simulation results for comparisons of outlier accommodation methods in order to provide recommendations for practice.

Methodology

In this study, a Monte Carlo program developed in the R programming language was conducted to simulate data, extract samples and calculate the statistics indices under a variety of conditions. First, three groups of univariate standard normal distribution data under different conditions were simulated by using the built-in R function rnorm. Samples of varied sample size and varied number of outliers were drawn from the same univariate normally-distributed data. For each condition, equal sample sizes were manipulated for three groups and a varied number of outliers were injected in only one group. ANOVA and Welch tests were performed using the same group of simulated data with both outliers included but with no treatment, and with outliers accommodated by the two types of Winsorization methods. Nonparametric tests were also performed using the same sample data with outliers included. For each condition, 10,000 replications were conducted. Type I error rate and statistical power for different outlier accommodation techniques and two different effect size conditions were computed and compared, and advantages and disadvantages of the outlier treatment techniques under different conditions noted.

Data Generation and Outlier Injection

The sample sizes (n = 20, 40, 60, 80, and 100) were manipulated in such a way that the three groups for statistical test always had equal sample sizes with the outlier(s) being inserted into only one group (group one). 200,000 normally distributed N(0, 1) cases were generated using the function rnorm. The generated population data were split into two data sets: data without outliers ($u - 3\sigma \leq x \leq u + 3\sigma$) and data with outliers ($x < u - 3\sigma$ and $x > u + 3\sigma$). Data for each sample were randomly selected from these two data sets. Previous research has investigated outlier impact on Type I error rate (Liao et al., 2016); this study repeated the Monte Carlo methods under the true null condition. Distinct from that effort, however, the performance of the Type I error rate by adding both one-end and two-end Winsorizing methods was considered in the current study. The mean for each group was 0. Additionally, two false null conditions were examined to display the performance of power rate under different treatment methods such as ANOVA, Welch, Nonparametric test and two types of Winsorizing. For the first false null condition, the means for group one, group two, and group three were set as 0.0, 0.3, and 0.6, respectively. For the second false null condition, 0.0, -0.3 and -0.6 were assigned respectively to the means of group one, two and three. The two conditions have equal magnitude of effect size.

Outliers were sampled from data beyond 3 standard deviations in both directions of the generated data, and the absolute value of outliers were injected into each sample; that is, all the inserted outliers are positive 1, 2, 3, 4 and 5 outliers and, for each sample size mentioned above, were investigated for both the Type I error analysis and power study under various treatment methods.

Monte Carlo Analysis

Under the true null hypothesis for each sample from the simulated population (e.g., $u_1 = 0$, $u_2 = 0$, $u_3 = 0$, $n = 40$, $n_{\text{outliers}} = 1, 2, 3, 4, 5$, sd = 1), two types of Winsorizing methods were used to examine the extent to which the inflation of the Type I error could be controlled: Winsorizing one end of data (the side with outliers) and Winsorizing both ends of data. The specified percentiles of Winsorizing for each condition are listed in Table 1, and the percentiles were performed through setting the value of parameter lambda (λ) in the R program. For example, when $N = 40$ and $n_{\text{outliers}} = 2$, $\lambda = 0.05$ (5th percentile) was employed. Under the conditions of one-end Winsorizing, only the outlier(s) were Winsorized; under the conditions of two-end Winsorizing, both the outlier(s) and the corresponding number of data on the opposite side were Winsorized.

Table 1. The percentile of Winsorizing (lambda λ)

	Number of outlier(s)				
Sample size	**1**	**2**	**3**	**4**	**5**
20	0.0500	0.1000	0.1500	0.2000	0.2500
40	0.0250	0.0500	0.0750	0.1000	0.1250
60	0.0200	0.0400	0.0500	0.0700	0.0900
80	0.0125	0.0250	0.0375	0.0500	0.0625
100	0.0100	0.0200	0.0300	0.0400	0.0500

Under the false null hypothesis, for each sample from the simulated population (e.g., $u_1 = 0$, $u_2 = 0.3$, $u_3 = 0.6$, $n = 20$, $n_{\text{outliers}} = 1, 2, 3, 4, 5$, sd = 1), ANOVA and Welch tests were used to explore the statistical power, that is, true rejection rates for the false null hypothesis. Statistical *p*-values were documented for data with no outliers, data with outliers, and data treated by two commonly-used outlier accommodation methods: nonparametric and Winsorizing.

Apart from the simulation procedures and data analyses, this study also adopted different verification approaches to validate data generation and collection. A small sample size (e.g., $N = 10$), small outlier number (e.g., 1 outlier), and small replications (e.g., 10 iterations) were carried out for generating dataset. Total rejection rates computed by hand were compared with the solution acquired from a cyber-program in order to manually verify data generation. A few normally- distributed sample data sets, simulated by the R program, were tested via the Statistical Package for the Social Sciences (SPSS) program. The data were confirmed to be indeed distributed normally. Various trials such as 1, 10, 100, and 1000 were employed for the stress-testing of R codes. All the results obtained from the specific R testing codes exhibited good performances under varied conditions.

Results

Simulation results are compiled in Table 2 and Table 3. The results include statistical power of parametric significance tests and different outlier accommodation techniques under two effect sizes (0, 0.3, 0.6; 0, -0.3, -0.6), five sample sizes (20, 40, 60, 80, and 100), and with six outlier conditions (outlier = 0, 1, 2, 3, 4, 5).

Table 2. Power of parametric significance tests and different outlier accommodation techniques under varied sample size, outlier conditions and effect size 0.0, 0.3, 0.6

Sample Size	Outlier	Parametric		Non-parametric	Wins.: one-end		Wins: two-end	
		ANOVA	Welch		ANOVA	Welch	ANOVA	Welch
N = 20	0	0.3720	0.3629	0.3473	0.3720	0.3629	0.3720	0.3629
	1	0.1555	0.1536	0.2250	0.2677	0.2619	0.2618	0.2573
	2	0.0663	0.0927	0.1312	0.1929	0.1939	0.1871	0.1879
	3	0.0451	0.0859	0.0831	0.1523	0.1589	0.1588	0.1578
	4	0.0573	0.0992	0.0653	0.1341	0.1421	0.1604	0.1614
	5	0.0989	0.1296	0.0720	0.1276	0.1378	0.2057	0.2042
N = 40	0	0.6723	0.6620	0.6419	0.6723	0.6620	0.6723	0.6620
	1	0.4966	0.4800	0.5425	0.5754	0.5623	0.5674	0.5561
	2	0.3313	0.3265	0.4393	0.4741	0.4648	0.4572	0.4500
	3	0.2148	0.2349	0.3493	0.3898	0.3859	0.3684	0.3655
	4	0.1365	0.1826	0.2712	0.3209	0.3249	0.2991	0.2980
	5	0.1003	0.1643	0.2127	0.2686	0.2785	0.2505	0.2561
N = 60	0	0.8507	0.8480	0.8245	0.8507	0.8480	0.8507	0.8480
	1	0.7520	0.7429	0.7699	0.7918	0.7852	0.7875	0.7815
	2	0.6340	0.6208	0.7070	0.7203	0.7143	0.7060	0.7025
	3	0.5068	0.5006	0.6380	0.6475	0.6397	0.6261	0.6201
	4	0.3817	0.3976	0.5605	0.5739	0.5673	0.5411	0.5378
	5	0.2818	0.3203	0.4877	0.5739	0.5673	0.4652	0.4677
N = 80	0	0.9386	0.9376	0.9235	0.9386	0.9376	0.9386	0.9376
	1	0.8940	0.8888	0.8958	0.9108	0.9081	0.9086	0.9054
	2	0.8301	0.8199	0.8652	0.8718	0.8689	0.8628	0.8611
	3	0.7449	0.7349	0.8211	0.8246	0.8197	0.8103	0.8056
	4	0.6509	0.6473	0.7748	0.7745	0.7681	0.7509	0.7469
	5	0.5483	0.5595	0.7232	0.7198	0.7133	0.6845	0.6830
N = 100	0	0.9748	0.9741	0.9658	0.9748	0.9741	0.9748	0.9741
	1	0.9575	0.9547	0.9536	0.9634	0.9619	0.9618	0.9608
	2	0.9253	0.9197	0.9373	0.9450	0.9431	0.9414	0.9392
	3	0.8811	0.8740	0.9180	0.9211	0.9162	0.9128	0.9088
	4	0.8234	0.8168	0.8955	0.8897	0.8863	0.8757	0.8729
	5	0.7561	0.7525	0.8630	0.8543	0.8511	0.8331	0.8305

Table 3. Power of parametric significance tests and different outlier accommodation techniques under varied sample size, outlier conditions and effect size 0.0, -0.3, -0.6

Sample Size	Outlier	Parametric		Non-parametric	Wins.: one-end		Wins: two-end	
		ANOVA	Welch		ANOVA	Welch	ANOVA	Welch
N = 20	0	0.3724	0.3640	0.3488	0.3724	0.3640	0.3724	0.3640
	1	0.4864	0.4405	0.4185	0.4564	0.4391	0.4994	0.4846
	2	0.6272	0.5462	0.4984	0.5549	0.5247	0.6414	0.6251
	3	0.7641	0.6590	0.5877	0.6471	0.6096	0.7680	0.7572
	4	0.8791	0.7709	0.6777	0.7253	0.6902	0.8670	0.8608
	5	0.9508	0.8712	0.7665	0.7925	0.7610	0.9314	0.9347
N = 40	0	0.6691	0.6621	0.6352	0.6691	0.6621	0.6691	0.6621
	1	0.7557	0.7354	0.6878	0.7343	0.7230	0.7515	0.7423
	2	0.8307	0.8015	0.7382	0.7924	0.7760	0.8241	0.8166
	3	0.8934	0.8613	0.7895	0.8402	0.8253	0.8842	0.8754
	4	0.9413	0.9081	0.8311	0.8835	0.8685	0.9273	0.9209
	5	0.9702	0.9456	0.8714	0.9148	0.9017	0.9587	0.9536
N = 60	0	0.8542	0.8472	0.8265	0.8542	0.8472	0.8542	0.8472
	1	0.9008	0.8909	0.8576	0.8902	0.8833	0.8979	0.8930
	2	0.9350	0.9245	0.8836	0.9189	0.9135	0.9296	0.9253
	3	0.9613	0.9495	0.9075	0.9411	0.9317	0.9555	0.9502
	4	0.9786	0.9677	0.9259	0.9588	0.9516	0.9727	0.9693
	5	0.9882	0.9804	0.9451	0.9706	0.9646	0.9837	0.9821
N = 80	0	0.9424	0.9398	0.9243	0.9424	0.9398	0.9424	0.9398
	1	0.9619	0.9585	0.9399	0.9581	0.9553	0.9606	0.9585
	2	0.9762	0.9723	0.9534	0.9695	0.9678	0.9738	0.9723
	3	0.9863	0.9824	0.9640	0.9793	0.9767	0.9843	0.9817
	4	0.9922	0.9896	0.9725	0.9862	0.9838	0.9905	0.9897
	5	0.9951	0.9940	0.9790	0.9910	0.9890	0.9941	0.9937
N = 100	0	0.9793	0.9790	0.9719	0.9793	0.9790	0.9793	0.9790
	1	0.9881	0.9872	0.9773	0.9861	0.9854	0.9872	0.9864
	2	0.9925	0.9907	0.9821	0.9903	0.9898	0.9915	0.9910
	3	0.9954	0.9943	0.9867	0.9933	0.9925	0.9943	0.9937
	4	0.9977	0.9968	0.9901	0.9954	0.9948	0.9969	0.9965
	5	0.9983	0.9977	0.9925	0.9968	0.9966	0.9979	0.9977

Outlier Impact on Power

Results for the first false null condition (mean = 0.0, 0.3, 0.6) are summarized in Figure 1 and Figure 2. As is shown in the figures, under the first false null condition, the presence of outliers caused significant decrease in the power of statistical testing. When sample size is as small as 20, with the presence of one

outlier, the power dropped by about 60% in ANOVA from 0.372 to 0.156. As sample size increases, the power decrease slowed down. When sample size is as large as 100, the power dropped only by less than 2% at the presence of an outlier.

Shown in Figure 2 is the statistical power when nonparametric and Winsorizing two-end methods were used under the first false null condition. From what is shown in the figure, outlier accommodation methods, though slightly different in effectiveness, can help diminish the impact of outlier on power. However, these outlier-robust measures can only diminish the impact but can hardly eliminate the impact.

Simulation results for the second false null condition (mean = 0.0, -0.3, -0.6) are summarized in Figure 3 and Figure 4. In contrast to the first null condition, under the second false null condition power was increased with the presence of outliers. The results further confirmed the impact of outliers on power rates, and indicated that, as the number of outliers increase, their impact on power increases as well.

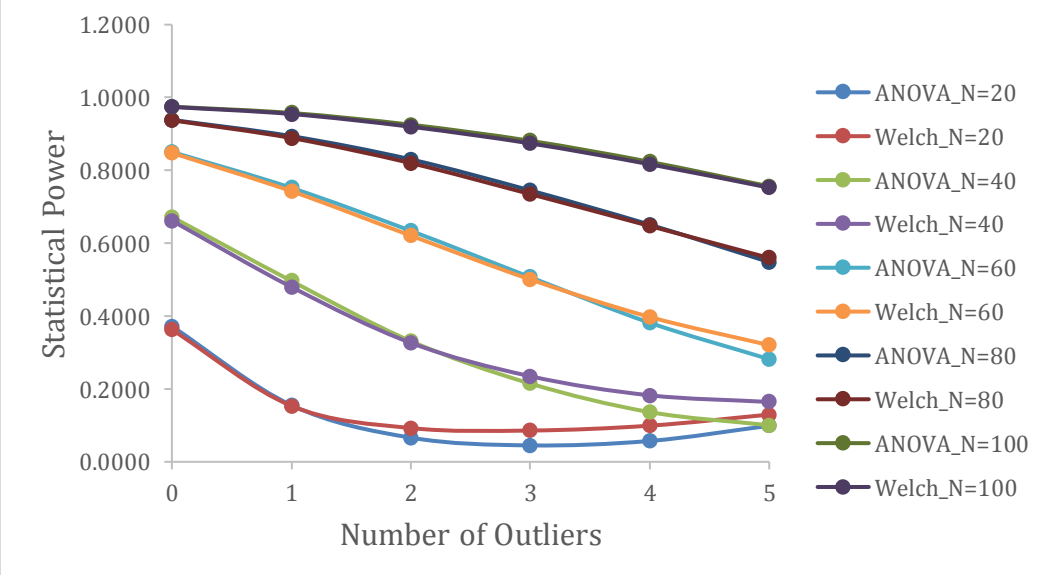


Figure 1. Statistical power for ANOVA and Welch with varied sample size and number of outliers when standardized group mean equals to 0.0, 0.3 and 0.6

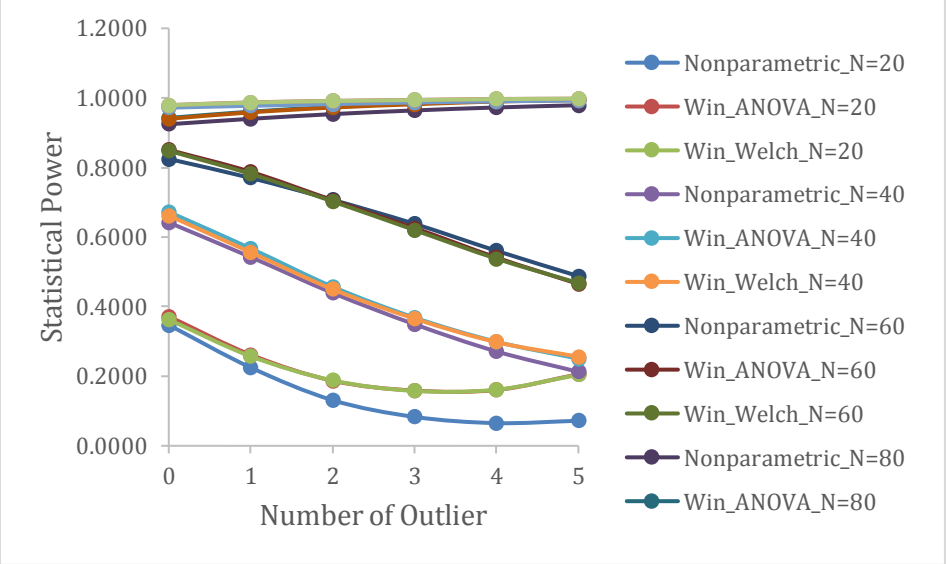


Figure 2. Statistical power for Nonparametric and Wisorizing two-end method with varied sample size and number of outliers when standardized group mean equals to 0.0, 0.3 and 0.6

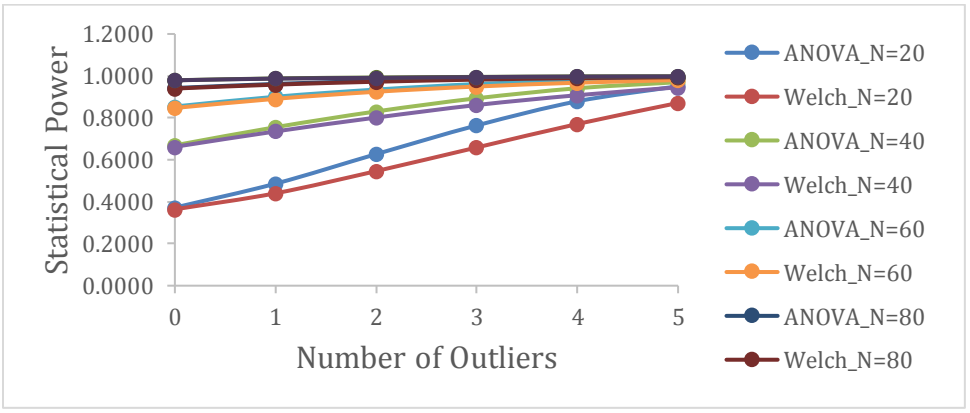


Figure 3. Statistical power for ANOVA and Welch with varied sample size and number of outliers when standardized group mean equals to 0.0, -0.3 and -0.6

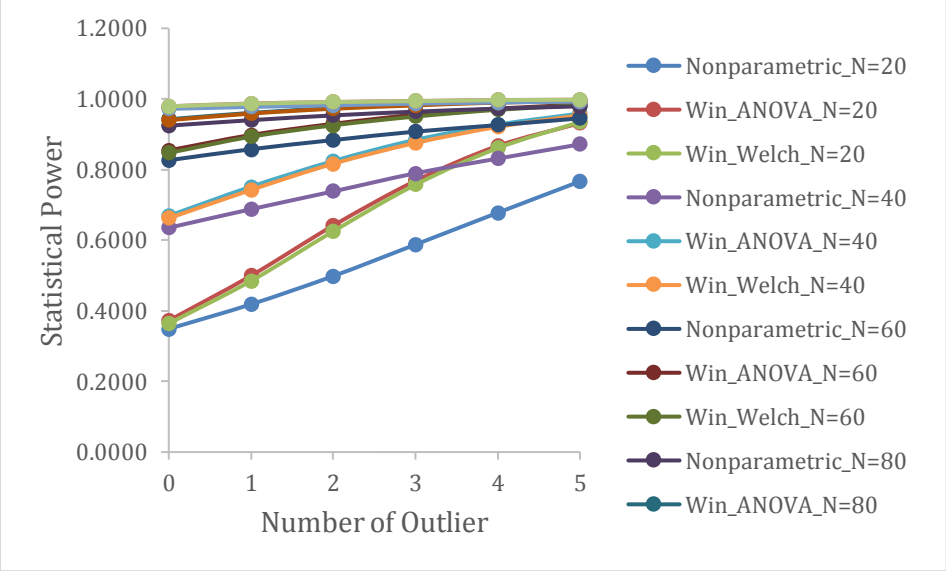


Figure 4. Statistical power for Nonparametric and Wisorizing two-end method with varied sample size and number of outliers when standardized group mean equals to 0.0, -0.3 and -0.6

Regarding the impact of outliers and effect size, in this study we inserted only positive outliers, and the results show that outlier impact is different for positive (mean = 0.0, 0.3, 0.6) and negative effect sizes (mean = 0.0, -0.3, -0.6).

Shown in Figure 5 is one of the examples of outlier impact on power with two effect sizes, and other results of other sample sizes showed similar trends. For positive effect size (0.0, 0.3, 0.6), the presence of outliers decreases power; for negative effect size (0.0, -0.3, -0.6), outliers increase power. Similar simulations with the effect size (-0.3, 0.0, 0.3) were conducted and yielded similar results as the effect size (0.0, 0.3, 0.6). Note that, in Figure 5: P1_ANOVA corresponds to Parametric ANOVA under the effect size 0.0, 0.3 and 0.6; P1_Welch corresponds to Parametric Welch under the effect size 0.0, 0.3 and 0.6; W11_ANOVA corresponds to Winsorizing one-end ANOVA under the effect size 0.0, 0.3 and 0.6; W11_Welch corresponds to Winsorizing one-end Welch under the effect size 0.0, 0.3 and 0.6; W12_ANOVA corresponds to Winsorizing two-end ANOVA under the effect size 0.0, 0.3 and 0.6; W12_Welch corresponds to Winsorizing two-end Welch under the effect size 0.0, 0.3 and 0.6; and P2_ANOVA corresponds to Parametric ANOVA under the effect size 0.0, -0.3 and -0.6.

Across all effect sizes, sample sizes, and numbers of outliers, ANOVA yields more power than Welch tests (see Table 2 and Table 3).

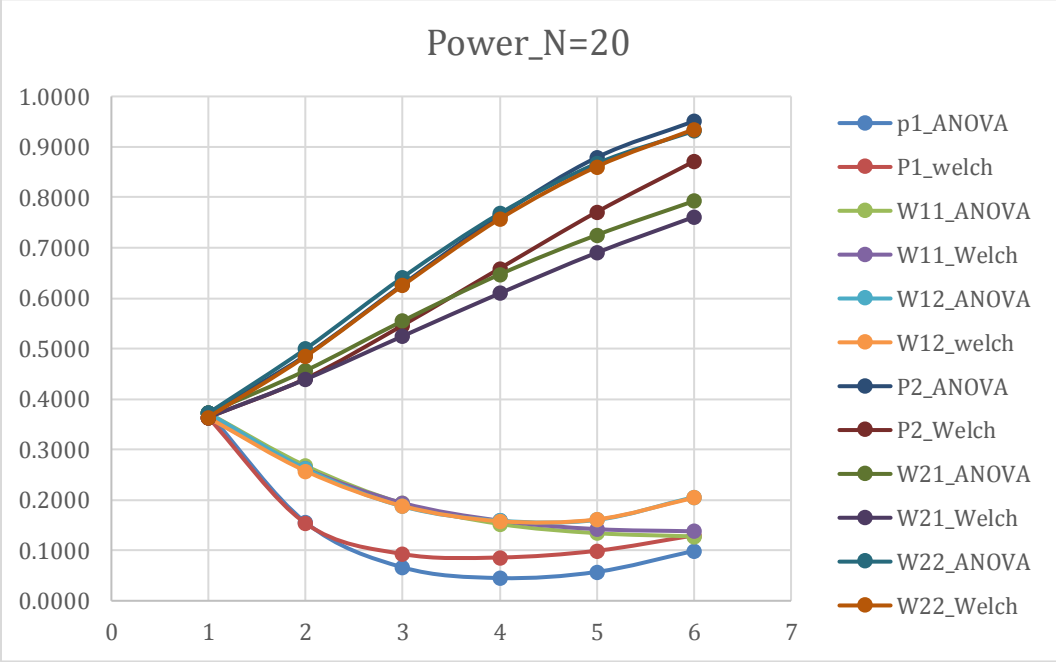


Figure 5. Statistical power for ANOVA and Welch with sample size = 20, number of outliers = 0, 1, 2, 3, 4, 5, and two effect sizes (standardized group mean equals to 0.0, 0.3, 0.6 and 0.0, -0.3, -0.6)

Comparison of Outlier Accommodation Methods on Power

Results on the effectiveness of the two outlier accommodation methods are now presented. As the results in Table 2 and Table 3 showed, outlier accommodation methods, including nonparametric tests and one-end and two-end Winsorizing, can help diminish the impact of outliers to a certain degree.

For the positive effect size, the decrease of power is a concern for parametric tests (ANOVA and Welch) when outliers are present. When sample sizes are small, the impact of outliers is stronger and outlier accommodation methods are relatively more effective in decreasing outlier impact; therefore they might be more useful in that case. For example, for positive effect size (0.0, 0.3, 0.6) and when $n = 40$ with 5 outliers, power decreased over 80%; outlier accommodation methods can increase the power by about 50%. Comparatively,

when sample sizes are large and when power decreases by about 50%, the outlier accommodation methods can increase the power by 10% at most.

Table 4. Type I error rates of nonparametric tests and Winsorizing method under varied sample sizes and outlier conditions

Sample Size	Outlier	Non-parametric	Winsorizing: one-end		Winsorizing: two-end	
			ANOVA	Welch	ANOVA	Welch
N = 20	0	0.0480	0.0492	0.0467	0.0492	0.0467
	1	0.0459	0.0522	0.0507	0.0615	0.0585
	2	0.0583	0.0707	0.0664	0.1043	0.1004
	3	0.0873	0.1034	0.0949	0.1940	0.1873
	4	0.1348	0.1563	0.1354	0.3339	0.3293
	5	0.2098	0.2282	0.1952	0.5053	0.5129
N = 40	0	0.0507	0.0528	0.0528	0.0528	0.0528
	1	0.0508	0.0556	0.0531	0.0597	0.0565
	2	0.0593	0.0674	0.0642	0.0856	0.0818
	3	0.0748	0.0898	0.0844	0.1288	0.1239
	4	0.0988	0.1247	0.1156	0.1950	0.1842
	5	0.1298	0.1709	0.1535	0.2878	0.2717
N = 60	0	0.0508	0.0497	0.0522	0.0497	0.0522
	1	0.0511	0.0517	0.0530	0.0546	0.0558
	2	0.0559	0.0617	0.0611	0.0713	0.0709
	3	0.0644	0.0776	0.0742	0.1015	0.0994
	4	0.0805	0.1052	0.0983	0.1461	0.1418
	5	0.0509	0.1399	0.1285	0.2087	0.1996
N = 80	0	0.0514	0.0546	0.0535	0.0546	0.0535
	1	0.0511	0.0543	0.0529	0.0566	0.0558
	2	0.0548	0.0621	0.0613	0.0694	0.0676
	3	0.0620	0.0770	0.0721	0.0931	0.0891
	4	0.0742	0.0990	0.0922	0.1315	0.1244
	5	0.0913	0.1303	0.1196	0.1787	0.1674
N = 100	0	0.0509	0.0489	0.0483	0.0489	0.0483
	1	0.0513	0.0506	0.0496	0.0527	0.0516
	2	0.0531	0.0551	0.0554	0.0609	0.0601
	3	0.0606	0.0685	0.0669	0.0832	0.0807
	4	0.0697	0.0879	0.0844	0.1121	0.1074
	5	0.0795	0.1116	0.1043	0.1503	0.1434

Regarding a comparison between nonparametric tests and Winsorizing, for the first false null condition with the effect size (0.0, 0.3, 0.6), Winsorizing

performed a little better in obtaining higher power than nonparametric test. In general, both nonparametric and Winsorizing show similar effects in increasing power. Similarly, a comparison of one-end and two-end Winsorizing methods shows that the two Winsorizing methods yield similar results, with one-end Winsorizing having slightly better performance in controlling outlier impact on power.

It is suggested by the simulation results and comparison of outlier accommodation methods above that, when examining the robustness and effectiveness of outlier accommodation methods, both power and Type I error should be taken into consideration. In our earlier study (Liao et al., 2016), we compared the effectiveness of nonparametric tests and one-end Winsorizing in controlling outlier impact on Type I error rates. In this study, based on earlier results, a comparison of Type I error rates with one-end and two-end Winsorizing was conducted. Table 4 is a summary of Type I error rates from previous studies with new results on the comparison of one-end and two-end Winsorizing methods. For effect size (0.0, 0.3, 0.6), although both nonparametric and Winsorizing show similar effects in increasing power, nonparametric methods yield the lowest Type I error rates across different sample sizes and numbers of outliers. For effect size (0.0, -0.3, -0.6), as the presence of outliers increases power, there is less concern regarding power but more regarding Type I error rate. Nonparametric tests were shown to be the most robust in controlling Type I error among all accommodation methods. Between one-end and two-end Winsorizing, one-end Winsorizing consistently performed better in controlling outlier impact on Type I error and power. In addition, one-end Winsorizing becomes more effective when the number of outliers gets bigger.

Conclusion

It was concluded previously that the impact of outliers on nonparametric tests in terms of Type I error rates alone depends on sample size and the number of outliers (Liao et al., 2016). When sample size is relatively large (e.g., n = 80 and 100), a nonparametric test has a good control of Type I error. When the sample size is small, there is non-ignorable inflation in Type I error caused by outlier influence, especially with two and more outliers present. Furthermore, it is the number of outliers that seems to matter when it comes to the issue of outlier impact on the statistical results, regardless of the sample size. No matter how large the sample size is, the false rejection rates almost adhere to the nominal significance level (0.05) when the number of outliers is less than two, indicating

that no accommodation techniques are necessary. As the number of outliers increases, the inflation of Type I errors begins to appear.

This simulation study built on the previous simulation stud. It further compared outlier accommodation with one-end and two-end Winsorizing and followed up with outlier impact on power to discuss outlier accommodation methods with consideration of both power and Type I error. This study has yielded new evidence regarding outlier impact on power, and the comprehensive effectiveness of the two commonly-used outlier accommodation methods in controlling outlier impact on Type I error and power.

First, the results show that the location of outliers could affect the direction of their impact. When only positive outliers were inserted, power decreases for positive effect size (mean = 0.0, 0.3, 0.6) and increases for negative effect size (mean = 0.0, -0.3, -0.6). Therefore, depending on the location of the outliers, the researcher needs to decide when outlier impact on power is a big concern.

Secondly, among parametric tests, ANOVA, and Welch tests yield similar results in the presence of outliers; Welch tests consistently have better control in Type I error rate. Winsorizing seems a little more effective compared with nonparametric tests in controlling outlier impact on power, but since the difference is less than 5% and nonparametic tests always have better control of Type I error inflation, the nonparametric tests seem the safest approach across most conditions.

Lastly, Winsorizing only one end seems better than both ends in controlling Type I error inflation and outlier impact on power. Therefore, it is recommended that when all outliers are on the same side, one-end Winsorizing is the most useful approach.

Both nonparametric and Winsorizing methods have similar effects in diminishing outlier impact on power, yet when deciding on an accommodation method, it is necessary to comprehensively consider both power and Type I error. Therefore, the nonparametric seems safest because the Type I error remains only a little inflated with more outliers but it generally has higher power.

Outliers will almost inevitable exist in educational datasets and, in practice, removing outliers is still a common approach (Bekker, 2014). It is therefore highly recommended to examine the reason for outlier occurrence and, if the reasons are obscure or cannot be traced, our recommendation is to retain the outlier and use appropriate outlier accommodation methods to minimize outlier impact in statistical testing.

Acknowledgments

This research was supported by School of English for Specific Purposes, Beijing Foreign Studies University, grant ZJ1513.

References

Allison, D. B, Gorman, B. S., & Primavera, L. H. (1993). Some of the most common questions asked of statistical consultants: Our favorite responses and recommended readings. *Genetic, Social, and General Psychology Monographs, 119*(2), 153-185.

Analytical Methods Committee. (1989). Robust statistics-how not to reject outliers: Part 1. Basic concepts. *Analyst*, *114*, 1693-1697. doi: 10.1039/an9891401693

Bakker, M., & Wicherts, J. M. (2014). Supplemental material for outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*. doi: 10.1037/met0000014.supp

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, UK: Wiley.

Berkane, M., & Bentler, P. M. (1988). Estimation of contamination parameters and identification of outliers in multivariate data. *Sociological Methods and Research, 17*(1), 55-64. doi: 10.1177/0049124188017001003

Chow, S.-M., Hamaker, E. L., & Allaire, J. C. (2009). Using innovative outliers to detect discrete shifts in dynamics in group-based state-space models. *Multivariate Behavioral Research, 44*(4), 465-496. doi: 10.1080/00273170903103324

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological), 48*(2), 133-169. Available from http://www.jstor.org/stable/2345711

Dixon, W. J., & Yuen, K. K. (1974). Trimming and Winsorization: A review. *Statistische Hefte, 15*(2), 157-170. doi: 10.1007/BF02922904

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). New York, NY: Norton.

Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations* (2nd ed.). New York, NY: Wiley.

Hampel, F. R. (2001). *Robust statistics: A brief introduction and overview*. Research Report No. 94. Zürich, Switzerland: Eidgenössische Technische Hochschule. Retrieved from ftp://ess.r-project.org/Research-Reports/94.pdf

Hawkins, D. M. (1974). The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association, 69*(346), 340-344. doi: 10.2307/2285654

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17-22. doi: 10.2307/2683469

Jarrell, M. G. (1991, November). *Multivariate outliers: Review of the literature*. Paper presented at the Annual Meeting of the Mid-South educational research Association, Lexington, KY.

Li, Y., An, Q., Wanich, W., Lewis, M., Huang, Y., & Brooks, G. (2009, October). *Type I error rates and power for multiple comparisons when extreme values exist in data*. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, St. Louis, MO.

Liao, H., Li, Y., & Brooks, G. (2016). Outlier impact and accommodation methods: Multiple comparisons of type I error rates. *Journal of Modern Applied Statistical Methods, 15*(1), 452-471. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol15/iss1/23

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*(3), 473-486. doi: 10.1111/j.1744-6570.1991.tb02401.x

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*(411), 633-639. doi: 10.2307/2289995

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*(2), 334-344. doi: 10.1037/0033-2909.95.2.334

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. New York, NY: Allyn & Bacon.

Warner R. M. (2012). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage.

Wilcox, R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314. doi: 10.1037/0003-066x.53.3.300

Wilcox, R. (2012). *Modern statistics for the social and behavioral sciences: A practical introduction*. Boca Raton, FL: CRC Press

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology, 121*(4), 391-401. doi: 10.1080/00221309.1994.9921213

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *The Journal of Experimental Education, 64*(1), 71-85. doi: 10.1080/00220973.1995.9943796

Zimmerman, D. W. & Zumbo, B. D. (1990). The relative power of the Wilcoxon-Mann-Whitney test and student t test under simple bounded transformations. *The Journal of General Psychology, 117*(4), 425-436. doi: 10.1080/00221309.1990.9921148

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 279-295.
doi: 10.22237/jmasm/1493597700

ISSN 1538 − 9472

A Note on Determination of Sample Size from the Perspective of Six Sigma Quality

Joghee Ravichandran
Amrita Vishwa Vidyapeetham, Amrita University
Coimbatore, India

In most empirical studies (clinical, network modeling, and survey-based and aeronautical studies, etc.), sample observations are drawn from population to analyze and draw inferences about the population. Such analysis is done with reference to a measurable quality characteristic of a product or process of interest. However, fixing a sample size is an important task that has to be decided by the experimenter. One of the means in deciding an appropriate sample size is the fixation of error limit and the associated confidence level. This implies that the analysis based on the sample used must guarantee the prefixed error and confidence level. Although there are methods to determine the sample size, the most commonly used method requires the known population standard deviation, the preset error and the confidence level. Nevertheless, such methods cannot be used when the population standard deviation is unknown. Because the sample size is to be determined, the experimenter has no clue to obtain an estimate of the unknown population standard deviation. A new approach is proposed to determine sample size using the population standard deviation estimated from the product or process specification from the perspective of Six Sigma quality with a goal of 3.4 defects per million opportunities (DPMO). The aspects of quality improvement through variance reduction are also presented. The method is effectively described for its use and is illustrated with examples.

Keywords: Coefficient of variation, DPMO, error, confidence level, sample size, Six Sigma quality, stopping criteria

Introduction

In most empirical studies, sample observations are often used to analyze and draw inferences about the population. Though a larger sample size results in better conclusions, the choice of sample size is very important for such studies. This is due to the fact that a larger sample size may require too much time, resources, and cost and, at the same time, a smaller sample size may lead to inaccurate inferential

Joghee Ravichandran is a Professor in the Department of Mathematics, Amrita School of Engineering. Email him at: aishwar2@rediffmail.com.

results. Therefore, in practice, before the choice of sample size, the aspects of time, resources, and cost have to be taken into consideration in addition to sufficient statistical power. An experimenter also prefers to fix a sample size without much compromise on the two types of errors. The problem of sample size determination is quite common in the research areas such as clinical trials (Ando et al., 2015), network modeling (Krivitsky & Kolaczyk, 2015), and aeronautical studies (Suárez-Warden, Rodriguez, Hendrichs, García-Lumbreras, & Mendívil, 2015).

In order to know how large a sample size must be fixed, a number of factors may be considered by both statisticians and researchers. Sometimes, it depends on the nature of study of interest. That is, the study may be survey-based to find out the proportion of something, or may be to estimate the population mean, standard deviation, correlation coefficient, regression coefficients, etc. So, given the nature of a study, "how to conclude if the sample size used is enough and is the right representation of the population?" is the most commonly raised question.

From a normal population whose mean is, say, μ and standard deviation is, say, σ, a number of samples may be collected, from which respective sample means, say ($\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_i, \ldots$) can be computed. The difference between each sample mean and population mean can be thought of as an error. However, in practice and due to various reasons, an experimenter selects randomly only one sample of size, say, n, and computes a sample mean, say $\bar{X}$. Then the difference $|\bar{X} - \mu|$ is treated as an absolute error. Apart from this, a $(1 - \alpha)100\%$ confidence interval for μ can be constructed by setting

$$\mathrm{P}\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le +z_{\alpha/2} \right) = (1-\alpha)100\%$$

Here α is the level of significance or the probability of Type-I error. Therefore, an experimenter always prefers to fix the sample size n such that the absolute error is kept at minimum, that is, $|\bar{X} - \mu| \le \varepsilon$, $\varepsilon > 0$ with maximum confidence that can result from maintaining minimum Type-I error probability. Clearly, $\varepsilon = z_{\alpha/2}\,\sigma/\sqrt{n}$ and hence $n = \left(z_{\alpha/2}\,\sigma/\varepsilon\right)^2$.

Since the population standard deviation σ is usually unknown and the sample standard deviation cannot be used as it needs the sample size n, there is a difficulty in determining the sample size n. In this paper, under the normality assumption, it is proposed to estimate the unknown population standard deviation from the specification of the quality characteristic that is under study from the

perspective of Six Sigma quality (SSQ), which can ensure only 3.4 defects per million opportunities (DPMO); refer to Ravichandran (2006). This estimated standard deviation is then used to determine the sample size.

Process and product specifications play a major role in ensuring the degree of quality of a process or product. It may be noted that a unit of a product is said to be defective if it fails to meet the preset specification limits of the quality characteristic that is critical-to-quality. Setijono (2010) has considered the case of matching the SSQ limits to specification limits in order to estimate customer dissatisfaction (not meeting specification) and delight (meeting specification) in a survey related study. A similar study was done by Ravichandran (2016) from the perspective of process/product specification to estimate DPMO and extremely good parts per million opportunities (EGPMO) for higher the better and lower the better quality characteristics. A process or product that meets the specification target is always said to be stable. However, the process/product mean may move away from the target over a period of time. In the context of Six Sigma, this has prompted the practitioners to allow a shift up to $\pm 1.5\sigma$ (Lucas, 2002) as it can still produce only 3.4 DPMO. It has been argued that, though such a shift from the target is not acceptable to many researchers due to lack of either theoretical or empirical justification (Antony, 2004), there is a strong belief among the Six Sigma practitioners that no process can maintain on its own target in the long run. Therefore, the population mean and standard deviation estimated using the proposed method are expected to satisfy the Six Sigma goal of 3.4 DPMO.

Sampling from Normal Population

Let the quality characteristic X follow a normal distribution with mean μ and variance σ^2. That is,

$$\mathrm{E}(X) = \mu \quad \text{and} \quad \mathrm{V}(X) = \sigma^2$$

Let $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ be a sample of size n drawn from this population. Then the sample mean $\bar{X}$ and sample variance S^2 are given as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 \tag{2}$$

It may be noted that the sample mean and variance given in (1) and (2) are the unbiased estimators of the mean μ and variance σ^2, respectively. That is,

$$\mathrm{E}\left(\bar{X}\right) = \mu \quad \text{and} \quad \mathrm{E}\left(S^2\right) = \sigma^2$$

Because the sample mean $\bar{X}$ itself can be thought of as a random variable as it can vary for varying samples, the mean and variance of the sample mean itself can be shown as μ and σ^2/n. It is a proven result that the sample mean $\bar{X}$ also follows the normal distribution with mean μ but with variance σ^2/n. In general, the standard deviation σ/n of the sample mean $\bar{X}$ is known as standard error (SE).

Standard Normal Distribution

It may be recalled that if the underlying distribution of the random variable X has mean μ and known variance σ^2, then we can define a standard normal variate, say Z, as

$$Z = \frac{X-\mu}{\sigma} \quad \text{or} \quad Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \tag{3}$$

or in general, equation (3) can be written as

$$Z = \frac{\text{sample statistic} - \mathrm{E}\left(\text{sample statistic}\right)}{\mathrm{SE}\left(\text{sample statistic}\right)}$$

which has mean 0 and variance 1. Here, E(*) represents expectation and SE(*) represents standard error. However, if the standard deviation σ is unknown then Z is observed to be not a standard normal variate. Under this circumstance, we replace the unknown standard deviation σ by the sample standard deviation given by S and construct a variable called Student's T as

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{4}$$

which follows the Student's T distribution with $n-1$ degrees of freedom. It may also be noted that the Student's T variable is defined when the sample size n is small.

Error and Sample Size

It may be noted that fixing the sample size n is a major concern in statistical inference problems. As discussed earlier, a large sample size, though preferred, may be expensive, laborious, and time-consuming, while a small sample may result in poor and inconsistent inferential decisions. Statistical errors – Type-I and Type-II errors – are also influenced by the size of the sample. Therefore, there needs to be a balance between these two types of error. It is preferable to choose n such that the *size*, say α, which is the probability of Type-I error and *power*, say $1-\beta$, where β is the probability of Type-II error, are optimum and vice-versa. Given the Type-I error probability α, it is known that

$$\begin{aligned} &\mathrm{P}\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le +z_{\alpha/2}\right) = (1-\alpha)100\% \\ &\Rightarrow \mathrm{P}\left(\left|\bar{X}-\mu\right| \le +\frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = (1-\alpha)100\% \end{aligned} \tag{5}$$

Here $\pm z_{\alpha/2}$ can be obtained by setting $\mathrm{P}(Z < -z_{\alpha/2}) = \mathrm{P}(Z > +z_{\alpha/2}) = \alpha/2$ with an assumed value of $\mu = \mu_0$ (null hypothesis is true), and hence $Z \sim \mathrm{N}(0, 1)$. Now it is supposed that an experimenter would like to have the difference (*error*) between the sample mean $\bar{X}$ and the unknown population mean μ to be less than or equal to a pre-specified negligible value, say $\varepsilon\,(>0)$, with the confidence level $(1-\alpha)100\%$. This implies that

$$\frac{\sigma}{\sqrt{n}} z_{\alpha/2} = \varepsilon \quad \Rightarrow \quad n = \left(\frac{\sigma}{\varepsilon} z_{\alpha/2}\right)^2 \tag{6}$$

(Refer to Montgomery & Runger, 2003; Ravichandran, 2010). One way of choosing ε is to allow the difference between $\bar{X}$ and μ as some $\delta\,(>0)$ percentage of μ, that is $\varepsilon = (\delta/100)\mu$. Therefore we have

$$n = \left(\frac{\sigma}{(\delta/100)\mu} z_{\alpha/2} \right)^2 = \left\{ \frac{1}{\delta} \left(\frac{\sigma}{\mu} 100 \right) z_{\alpha/2} \right\}^2 \tag{7}$$

Accordingly, if μ and σ are known, then for the known values of $(\sigma/\mu)100$ (note that $(\sigma/\mu)100$ gives the coefficient of variation (CV)) and for different δ values, the sample size can be determined by fixing α values. Table 1 shows such sample size values for

(i) CV = $(\sigma/\mu)100$ = 2.5% (2.5) 20%
(ii) δ = 1.0, 2.5, 50
(iii) α = 0.01, 0.05, 0.10

Readers may note that, in Table 1, CV = x% and $\delta = y$ means $x = \sigma/\mu$ and $y = \delta/100$ so that

$$n = \left\{ \frac{x}{y} z_{\alpha/2} \right\}^2$$

Table 1 can now readily be used by the experimenters for sampling or can be used as a guideline for determining sample size for other combinations of parameters. If both Type-I and Type-II error probabilities are known, then the sample size n given in equation (6) can also be written as

$$n \approx \left(\frac{\sigma}{\varepsilon} \left[z_{\alpha/2} + z_{\beta} \right] \right)^2 \tag{8}$$

Here z_β can be obtained by setting β equal to $\mathrm{P}\left(Z < +z_{\alpha/2} - \varepsilon\sqrt{n}/\sigma\right) - \mathrm{P}\left(Z < -z_{\alpha/2} - \varepsilon\sqrt{n}/\sigma\right)$ with an assumed mean value $\mu = \mu_1 = \varepsilon\sqrt{n}/\sigma$ (alternative hypothesis is true) and hence $Z \sim \mathrm{N}\left(\varepsilon\sqrt{n}/\sigma, 1\right)$. It is observed that the approximation in (8) holds good if $\mathrm{P}\left(Z \le -z_{\alpha/2} - \varepsilon\sqrt{n}/\sigma\right)$ is small (= 0) compared to β for the sample size given in (8). Refer to Montgomery and Runger (2003) for more details. Therefore, $\mathrm{P}(Z < -z_\beta) = \beta$ implies that $-z_\beta = +z_{\alpha/2} - \varepsilon\sqrt{n}/\sigma$. Following (7), we have

Table 1. Sample size values according to equation (7)

		Size (α)		
CV	**δ**	**0.01**	**0.05**	**0.10**
2.5	1.0	42	24	17
	2.5	7	4	3
	5.0	2	1	1
5.0	1.0	166	96	68
	2.5	27	15	11
	5.0	7	4	3
7.5	1.0	374	216	153
	2.5	60	35	25
	5.0	15	9	6
10.0	1.0	666	384	272
	2.5	107	61	44
	5.0	27	15	11
12.5	1.0	1040	600	425
	2.5	166	96	42
	5.0	42	24	17
15.0	1.0	1498	864	613
	2.5	240	138	98
	5.0	60	35	25
17.5	1.0	2039	1176	834
	2.5	326	133	82
	5.0	82	47	33
20.0	1.0	2663	1537	1089
	2.5	426	246	174
	5.0	107	61	44

$$n \approx \left\{ \frac{1}{\delta} \left(\frac{\sigma}{\mu} 100 \right) \left[z_{\alpha/2} + z_{\beta} \right] \right\}^2 \tag{9}$$

From Table 1, the following observations can easily be made:

(i) For a fixed CV, as the error δ increases, the sample size n decreases meaning that smaller sample size will result in higher error and vice-versa.

(ii) For a fixed CV, as the Type-I error probability α increases, the sample size n decreases meaning that smaller sample size will result in higher degree of Type-I error probability (size) and vice-versa.

(iii) As CV increases, the sample size increases and vice-versa. This means that if the CV is less, then fewer sample observations are sufficient to achieve the error levels.

If μ is zero, then it is always wise to use the formula involving ε given in equation (6) rather than using the formula involving $\delta\mu$ given in (7). Values for ε can be assumed to be 10^{-2}, 10^{-3}, 10^{-4}, etc. If σ is unknown, S cannot be used in (6) or (7) since $E(S) \neq \sigma$. But, though $E(S/c_4) = \sigma$ where c_4 is an appropriate constant, one cannot use c_4 and S since both of them depend on sample size n. Therefore, using

$$n = \left(\frac{S/c_4}{\varepsilon} t_{v,\alpha/2} \right)^2 \quad \text{or} \quad n = \left(\frac{S/c_4}{(\delta/100)\mu} t_{v,\alpha/2} \right)^2 \tag{10}$$

respectively, as replacement of (6) or (7) for sample size determination is erroneous.

Stopping Criteria in Simulations

There are situations, such as simulations, where it is important to decide when to stop the simulation. Under these circumstances, the simulations are run for a preset number n_1 of times (i.e., sample of size n_1) and then the sample mean $\bar{X}_1$, standard deviation S_1, c_4^1, and $t_{v_1,\alpha/2}$ are computed for the quality characteristic of interest, say X. The simulation is stopped if the following condition is satisfied (refer to Yeap, 1998):

$$n_1 \geq \left(\frac{S_1/c_4^1}{\varepsilon} t_{v_1,\alpha/2} \right)^2 \quad \text{or} \quad n_1 \geq \left(\frac{S_1/c_4^1}{(\delta/100)\bar{X}_1} t_{v_1,\alpha/2} \right)^2$$

Otherwise, collect the next observation from the next simulation so that $n_2 = n_1 + 1$, from which $\bar{X}_2$, S_2, c_4^2, and $t_{v_2,\alpha/2}$ are computed to verify if

$$n_2 \geq \left(\frac{S_2 / c_4^2}{\varepsilon} t_{\nu_2, \alpha/2} \right)^2 \quad \text{or} \quad n_2 \geq \left(\frac{S_2 / c_4^2}{(\delta/100)\bar{X}_2} t_{\nu_2, \alpha/2} \right)^2$$

In general, the simulation is stopped after n_i, $i = 1, 2, \ldots$, simulations if

$$n_i \geq \left(\frac{S_i / c_4^i}{\varepsilon} t_{\nu_i, \alpha/2} \right)^2 \quad \text{or} \quad n_i \geq \left(\frac{S_i / c_4^i}{(\delta/100)\bar{X}_i} t_{\nu_i, \alpha/2} \right)^2, \quad i = 1, 2, \ldots$$

where the mean $\bar{X}_i$, standard deviation S_i, c_4^i, and $t_{\nu_i, \alpha/2}$ are computed form the sample of size n_i, that is after i simulations.

A method is proposed here to estimate the unknown population standard deviation σ from the perspective of the concept of SSQ. This sample size can ensure the conformance of the process to the Six Sigma goal of 3.4 DPMO.

Sample Size Determination based on Six Sigma Quality

Consider a measurable quality characteristic, say X, that follows normal process with mean $T = \mu$ and variance σ^2. Because not all values of X towards the tails of the distribution are acceptable, the specification of X is usually given in the form $T \pm K\sigma$, where T is the target or population mean, K is a positive constant, and σ is the population standard deviation. Notationally, $X \sim \mathrm{N}(T, \sigma^2)$ and $\mathrm{P}(T - K\sigma \leq X \leq T + K\sigma) = 1 - \alpha_K$, where α_K is a prespecified probability value such that $\alpha_K = \mathrm{P}(X < T - K\sigma) + \mathrm{P}(X > T + K\sigma)$. From $T \pm K\sigma$, we get half of the process spread as $K\sigma = d$ (say) (also refer to Lin, 2006), which implies $\sigma = d/K$ and hence we have $\hat{\sigma}_{\mathrm{SS}} = \sigma = d/K$. Therefore, we have $\hat{\sigma}_{\mathrm{SS}} / \sqrt{n_{\mathrm{SS}}} = (d/K) / \sqrt{n_{\mathrm{SS}}}$. Now equation (5) becomes

$$\Rightarrow \mathrm{P}\left(\left| \bar{X} - \mu \right| \leq + \frac{d/K}{\sqrt{n_{\mathrm{SS}}}} z_{\alpha_K/2} \right) = (1 - \alpha_K)100\% \tag{11}$$

and hence equations (6) and (7) become, respectively:

$$n_{\mathrm{SS}} = \left(\frac{d/K}{\varepsilon} z_{\alpha_K/2} \right)^2 \tag{12}$$

Table 2. Determination of α_K and $z_{\alpha_K/2}$

K	DPMO	α_K	$z_{\alpha_K/2}$
3.0	66810.63	0.1336210	1.50
3.5	22750.35	0.0455010	2.00
4.0	6209.70	0.1241900	2.50
4.5	1349.97	0.0027000	3.00
5.0	232.67	0.0004650	3.50
5.5	31.69	0.0000634	4.00
6.0	3.40	0.0000068	4.50

and

$$n_{\mathrm{SS}} = \left(\frac{d/K}{(\delta/100)\,\mu} z_{\alpha_K/2} \right)^2 \tag{13}$$

Here, K represents the current sigma quality level (SQL) of the process. For example, if $K = 6$, then we have DPMO = 3.4 either on left tail or on right tail. Therefore, $\alpha_K = 6.8 \times 10^{-6}$ implies $z_{\alpha_K/2} = 4.50$. In (13), if μ is unknown, then the same can be replaced by the specification target T.

The computation of the values of $z_{\alpha_K/2}$ with different SQLs is discussed as follows: If the process is operating at a Three Sigma level, then we have the current quality level as $K = 3$. It may be noted that, with allowable shift, a Three Sigma process may result in 66810.63 DPMO. Once this level is maintained, and if there is a scope for improvement, the practitioner may change the value of $z_{\alpha_K/2}$. Various DPMOs and the corresponding $z_{\alpha_K/2}$ values are given as shown in Table 2 (Harry, 1998; Lucas, 2002). Therefore, for SSQ process with 3.4 DPMO, (12) and (13) respectively become

$$n_{\mathrm{SS}} = \left(\frac{d/6}{\varepsilon} (4.50) \right)^2 \tag{14}$$

and

$$n_{SS} = \left(\frac{d/6}{(\delta/100)\mu}(4.50) \right)^2 \quad \text{or} \quad n_{SS} = \left(\frac{1}{\delta}\left(\frac{d/6}{\mu}100 \right)4.50 \right)^2$$
$$\Rightarrow n_{SS} = \left(\frac{1}{\delta}\left(\frac{\hat{\sigma}_{SS}}{\mu}100 \right)4.50 \right)^2 \tag{15}$$

Shown in Table 3 are sample size values from the perspective of Three Sigma (3σ), Four Sigma (4σ), Five Sigma (5σ), and Six Sigma (6σ) qualities for the following parameter set up:

(i) $CV_{SS} = (\hat{\sigma}_{SS}/\mu)x100 = 1.0, 2.5\%(2.5)20\%$
(ii) δ = 1.0, 2.5, 5.0
(iii) α_K = 0.1336210, 0.1241900, 0.0004650, 0.0000068

From Table 3, it can be seen that:

(i) For a fixed CV_{SS}, as the error δ increases, the sample size n decreases meaning that a smaller sample size will result in higher error and vice-versa.
(ii) For a fixed CV_{SS}, as the sigma quality decreases (that is, as the Type-I error probability α increases), the sample size n decreases meaning that a smaller sample size will result in poor sigma quality and vice-versa.
(iii) As CV increases the sample size increases and vice-versa. This means that if the CV is less, then fewer sample observations are sufficient to achieve the goal of SSQ of 3.4 DPMO. For example, if $(d/6)/\mu$ = 0.01 and the error percentage is δ = 1% of μ, then an inspection of a sample with 20 observations is sufficient to show if the process is meeting the Six Sigma goal of 3.4 DPMO.

Table 3 is an indicative one, and experimenters can use it as a guideline for determining the sample size for different parameter combinations. Looking at Tables 1 and 3, the values of

$$CV = \frac{\sigma}{\mu}100\% \quad \text{and} \quad CV_{SS} = \frac{\hat{\sigma}_{SS}}{\mu}100\%$$

are assumed as same for comparison purpose. However, in practice, the variation indicated by $\hat{\sigma}_{SS}$ in the case of a Six Sigma process is usually far below the normal process whose variation is indicated by σ. Therefore, reduced variation in Six Sigma may result in a good reduction in the sample size. See example 2 in the following section.

Table 3. Sample size n_{SS} for Six Sigma quality

CV_{SS}	δ	6σ	5σ	4σ	3σ
1	1.0	20	12	6	2
	2.5	3	2	1	-
	5.0	1	-	-	-
2.5	1.0	127	77	39	14
	2.5	20	12	6	2
	5.0	5	3	2	1
5.0	1.0	506	306	156	56
	2.5	81	49	25	9
	5.0	20	12	6	2
7.5	1.0	1139	689	352	127
	2.5	182	110	56	20
	5.0	46	28	14	5
10.0	1.0	2025	1225	625	225
	2.5	324	196	100	36
	5.0	81	49	25	9
12.5	1.0	3164	1914	977	352
	2.5	506	306	156	56
	5.0	127	77	39	14
15.0	1.0	4556	2756	1406	506
	2.5	729	441	225	81
	5.0	182	110	56	20
17.5	1.0	6202	3752	1914	689
	2.5	992	600	306	110
	5.0	248	150	77	28
20.0	1.0	8100	4900	2500	900
	2.5	1296	784	400	144
	5.0	324	196	100	36

Numerical Examples

Example 1

Yeap (1998) has given an example that the standard deviation of power samples measured from a circuit has been observed to have $\pm$ 20% fluctuations from the mean. Now the number of sample units (sample size) required to ensure that the experimenter is 99% confidence that the error of the sample mean is within $\pm$ 5% can be obtained by setting:

$$\sigma = 20\%\mu \Rightarrow \sigma/\mu = 0.2, \quad \delta = 5\% = 5/100$$

which, according to (7), gives

$$n = \left\{\frac{1}{\delta}\left(\frac{\sigma}{\mu}100\right)z_{\alpha/2}\right\}^2 = \left\{\frac{1}{5}(0.2)(100)(2.58)\right\}^2 \approx 107$$

However, for the SSQ requirement of 3.4 DPMO, the sample size can be obtained as

$$n_{SS} = \left\{\frac{1}{\delta}\left(\frac{\hat{\sigma}_{SS}}{\mu}100\right)z_{\alpha_K/2}\right\}^2 = \left\{\frac{1}{5}(0.2)(100)(4.50)\right\}^2 \approx 324$$

It is alarming to note that the SSQ process requires more sample observations in this example. This is due to the fact that the CV% is too high with $\sigma = 20\%\mu$, which is beyond expectation. However, it is presented here for illustration purpose to show that given this CV% and the specification of the quality characteristic of interest, it may require 324 sample observations to ensure that it is a Six Sigma process.

Example 2

Montgomery and Runger (2003) presented an example of vane-manufacturing process. The specifications on vane opening are given as 0.5030 $\pm$ 0.0010 inches. Let us suppose that we would like to draw a sample of size *n* so that the process average can lie around $\pm$ 0.05% of the target. Then the sample size meeting the SSQ requirement of 3.4 DPMO can be obtained by setting:

$$\mu = 0.5030, \quad \hat{\sigma}_{SS} = d/6 = 0.0010/6 = 0.000167,$$
$$\sigma/\mu = 0.000361, \quad \delta = 0.05\% = 0.05/100$$

This, according to (15), gives

$$n_{SS} = \left\{ \frac{1}{\delta} \left(\frac{\hat{\sigma}_{SS}}{\mu} 100 \right) z_{\alpha_K/2} \right\}^2 = \left\{ \frac{1}{0.05} (0.0361)(4.50) \right\}^2 \approx 11$$

If it is assumed that by past experience the standard deviation of this process is known as 0.00025, then the required sample size can be obtained by setting:

$$\mu = 0.5030, \sigma = 0.00025, \sigma/\mu = 0.000497, \delta = 0.05\% = 0.05/100$$

$$n = \left\{ \frac{1}{\delta} \left(\frac{\sigma}{\mu} 100 \right) z_{\alpha/2} \right\}^2 = \left\{ \frac{1}{0.05} (0.000497)(100)(4.50) \right\}^2 \approx 20$$

It may be noted that since $\sigma = 0.00025$, the process is at the level of 4σ only with $K = 4$ (that is, $4\sigma = (4)(0.00025) = 0.0010 = d$) and hence it requires more sample observations. Therefore, the process variation needs to be improved (reduced variation) with regard to standard deviation from $\sigma = 0.00025$ to $\sigma = 0.000167$ so that the process becomes a Six Sigma process with 3.4 DPMO.

If an experimenter is interested in drawing a sample of size n so that it meets the Four Sigma requirement of 6209.70 DPMO, then it can be obtained by setting:

$$\mu = 0.5030, \quad \hat{\sigma}_{SS} = d/4 = 0.0010/4 = 0.00025,$$
$$\hat{\sigma}_{SS}/\mu = 0.000497, \quad \delta = 0.05\% = 0.05/100$$

$$n_{SS} = \left\{ \frac{1}{\delta} \left(\frac{\hat{\sigma}_{SS}}{\mu} 100 \right) z_{\alpha_K/2} \right\}^2 = \left\{ \frac{1}{0.05} (0.000497)(100)(2.50) \right\}^2 \approx 6$$

Given the process conditions, it may be noted that a meager sample of size 6 is sufficient to meet the error constraints under Four Sigma quality of 6209.70 DPMO.

Example 3

Consider an example of a manufacturing process of a product in which the specification for the dimension of the product is set as 20 ± 6. For laboratory testing purposes it is proposed to collect sample units of the product. The error limit between sample mean and the target is set as $\pm 5\%$ of the target. Then the sample size meeting the SSQ requirement of 3.4 DPMO can be obtained by setting:

$$\mu = 20, \hat{\sigma}_{SS} = d/6 = 6/6 = 1, \hat{\sigma}_{SS}/\mu = 0.005, \delta = 5\% = 5/100$$

This, according to (15), gives

$$n_{SS} = \left\{ \frac{1}{\delta} \left(\frac{\hat{\sigma}_{SS}}{\mu} 100 \right) z_{\alpha_K/2} \right\}^2 = \left\{ \frac{1}{5} (0.05)(100)(4.50) \right\}^2 \approx 20$$

This can also be verified from Table 3. Now, after drawing a sample of size 20, the sample standard deviation is computed as 3.63, which is an indication that the process is only at an SQL of 6/3.63 = 1.65 sigma. Therefore, the process variation needs to be improved (reduced variation) with regard to standard deviation from 3.63 to 1 so that the process becomes a Six Sigma process with 3.4 DPMO.

Discussions and Conclusions

In this paper, first a discussion on the existing methods of sample size determination is presented. It is observed that such methods critically need the known population standard deviation. Therefore, a new approach is then presented that uses an estimate of population standard deviation from the perspective of the Six Sigma goal of 3.4 DPMO. The proposed method helps the experimenter to fix the sample size in such a way that the process either meets the SSQ requirement of 3.4 DPMO or can be improved towards the goal. This can be achieved by comparing the estimated standard deviation from the perspective of Six Sigma and the actual process standard deviation obtained after fixing the sample size. If the difference is wide, then we recommend using the stopping criteria approach by adding more samples until the requirements are met.

The proposed sample size determination method is studied and evaluated numerically. It is observed that as the CV% increases, the method recommends a

larger sample size to cover up the higher standard deviation and vice-versa. The approach is also demonstrated using suitable examples. In these examples, it is discussed that the proposed method not only helps in determining the sample size, it also prompts the experimenter to look for improvement opportunities, such as variance reduction exercises through quality improvement programs. As a future study, the case of proportions instead of measurable quality characteristic will be considered. Also, it will be attempted to propose a method for determining a sample of specific size from a finite size population.

References

Ando, Y., Hamasaki, T., Evans, S. R., Asakura, K., Sugimoto, T., Sozu, T., & Ohno, Y. (2015). Sample size considerations in clinical trials when comparing two interventions using multiple co-primary binary relative risk contrasts. *Statistics in Biopharmaceutical Research, 7*(2), 81-94. doi: 10.1080/19466315.2015.1006373

Antony, J. (2004). Some pros and cons of Six Sigma: An academic perspective. *The TQM Magazine, 15*(4), 303-306. doi: 10.1108/09544780410541945

Harry, M. J. (1998). Six Sigma: A breakthrough strategy for profitability. *Quality Progress, 31*(5), 60-64.

Krivitsky, P. N., & Kolaczyk, E. D. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statistical Science, 30*(2), 184-198. doi: 10.1214/14-sts502

Lin, G-H. (2006). A decision-making procedure on process centering: A lower bound of the estimated accuracy index. *Journal of Statistics and Management Systems, 9*(1), 205-224. doi: 10.1080/09720510.2006.10701203

Lucas, J. M. (2002). The essential Six Sigma. *Quality Progress, 35*(1), 27-31.

Montgomery, D. C., & Runger, G. C. (2003). *Applied statistics and probability for engineers* (3rd ed.). Hoboken, NJ: John Wiley and Sons, Inc.

Ravichandran, J. (2006). Setting up a quality specification. *Six Sigma Quality Magazine, 5*(2), 26-30. Retrieved from http://asq.org/pub/sixsigma/past/volume5-issue2/ssfmv5i2ravichandran.pdf

Ravichandran, J. (2010). *Probability and statistics for engineers*. Delhi, India: Wiley India.

Ravichandran, J. (2016). Estimation of DPMO and EGPMO for higher-the-better and lower-the-better quality characteristics for quality evaluation. *Total Quality Management & Business Excellence, 27*(9-10), 1112-1120. doi: 10.1080/14783363.2015.1060852

Setijono, D. (2010). Normal approximation through data replication when estimating DisPMO, DePMO, left-side and right-side Sigma levels from non-normal data. *International Journal of Quality & Reliability Management, 27*(3), 318-332. doi: 10.1108/02656711011023302

Suárez-Warden, F., Rodriguez, M., Hendrichs, N., García-Lumbreras, S., & Mendívil, E. G. (2015). Small sample size for test of training time by augmented reality: An aeronautical case. *Procedia Computer Science, 75*, 17-27. doi: 10.1016/j.procs.2015.12.190

Yeap, G. K. (1998). *Practical low power digital VLSI design*. Boston, MA: Kluwer Academic Publishers. doi: 10.1007/978-1-4615-6065-4

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 296-307.
doi: 10.22237/jmasm/1493597760

ISSN 1538 − 9472

An Extended Weighted Exponential Distribution

Abbas Mahdavi
Vali-e-Asr University of Rafsanjan
Rafsanjan, Iran

Leila Jabbari
Vali-e-Asr University of Rafsanjan
Rafsanjan, Iran

A new class of weighted distributions is proposed by incorporating an extended exponential distribution in Azzalini's (1985) method. Several statistics and reliability properties of this new class of distribution are obtained. Maximum likelihood estimators of the unknown parameters cannot be obtained in explicit forms; they have to be obtained by solving some numerical methods. Two data sets are analyzed for illustrative purposes, and show that the proposed model can be used effectively in analyzing real data.

Keywords: Exponential distribution, extended exponential distribution, hazard rate function, maximum likelihood estimation, weighted exponential distribution

Introduction

Adding an extra parameter to an existing family of distribution functions is common in statistical distribution theory. Introducing an extra parameter often brings more flexibility to a class of distribution functions, and it can be very useful for data analysis purposes. Azzalini (1985) introduced the skew normal distribution by introducing an extra parameter to bring more flexibility to the normal distribution. Afterwards, extensive works on introducing shape parameters for other symmetric distributions have been defined, and several properties and their inference procedures have been discussed by several authors; see for example Balakrishnan and Ambagaspitiya (1994), Arnold and Beaver (2000), and Nadarajah (2009).

Recently, there has been an attempt to use Azzalini's method for non-symmetric distributions. Gupta and Kundu (2009) introduced a class of weighted exponential (WE) distribution that has a shape parameter. It is said that a random variable X follows the WE(α, λ) distribution if its density function is given by

Abbas Mahdavi is Faculty of Mathematical Sciences in the Department of Statistics. Email them at: a.mahdavi@vru.ac.ir. Leila Jabbari is Faculty of Mathematical Sciences in the Department of Statistics.

$$\mathrm{f}\left(x;\alpha,\lambda\right)=\frac{\alpha+1}{\alpha}\lambda\mathrm{e}^{-\lambda x}\left(1-\mathrm{e}^{-\alpha\lambda x}\right) \tag{1}$$

where $x > 0$, $\alpha > 0$, and $\lambda > 0$.

Shakhatreh (2012) generalized the WE distribution to the two-parameter weighted exponential (TWE) distribution. A random variable X is said to have a TWE distribution with shape parameters $\alpha_1 > 0$, $\alpha_2 > 0$ and scale parameter $\lambda > 0$ if the PDF of X is given by

$$\mathrm{f}\left(x;\lambda,\alpha_1,\alpha_2\right)=\mathrm{k}\left(\alpha_1,\alpha_2\right)\lambda\mathrm{e}^{-\lambda x}\left(1-\mathrm{e}^{-\lambda\alpha_1 x}\right)\left(1-\mathrm{e}^{-\lambda\alpha_2 x}\right),\quad x>0 \tag{2}$$

where $\mathrm{k}\left(\alpha_1,\alpha_2\right)=\dfrac{\left(1+\alpha_1\right)\left(1+\alpha_2\right)\left(1+\alpha_1+\alpha_2\right)}{\alpha_1\alpha_2\left(2+\alpha_1+\alpha_2\right)}$.

It is observed that the WE and TWE distributions can provide a better fit for survival time data relative to other common distributions such as the gamma, Weibull, or generalized exponential distributions.

The aim of this study is to introduce an extended weighted exponential (EWE) distribution based on extended exponential (EE) distribution introduced by Gómez, Bolfarine, and Gómez (2014). A random variable X follows the EE distribution with parameters λ and β if its density function is given by

$$\mathrm{f}\left(x;\lambda,\beta\right)=\frac{\lambda^2\left(1+\beta x\right)e^{-\lambda x}}{\lambda+\beta},\quad x>0 \tag{3}$$

where $\lambda > 0$ and $\beta > 0$ with the notation $X \sim \mathrm{EE}(\lambda, \beta)$.

One of the goals of the introduction of the EWE is that involves the WE as its sub-model. The EWE has three parameters, one scale parameter and two shape parameters, which makes it more flexible in describing different types of real data than its sub model.

It is observed that the EWE distribution has several desirable properties. The generation of random samples from the EWE is straight forward. The maximum likelihood estimators (MLEs) of unknown parameters can be obtained by solving three nonlinear equations. For illustrative purposes we have analyzed the two real data sets. After analyzed the data using the EWE model, observe that EWE provides a better fit than the WE model and TWE.

Definition, Interpretations, and Generation

Definition: A random variable X is said to have an extended weighted exponential distribution with shape parameters $\alpha > 0$, $\beta > 0$ and scale parameter $\lambda > 0$, denoted by EWE(α, β, λ), if the density function of X is given as below

$$\mathrm{f}\left(x;\alpha,\beta,\lambda\right)=\frac{\left(1+\alpha\right)^{2}\lambda}{\alpha\left(\lambda\left(1+\alpha\right)+\alpha\beta\right)}\mathrm{e}^{-\lambda x}\left(\lambda+\beta-\left(\lambda+\beta+\alpha\beta\lambda x\right)\mathrm{e}^{-\alpha\lambda x}\right) \qquad (4)$$

for $x > 0$ and 0 otherwise.

Plots of the EWE density function for fixed scale parameter $\lambda = 1$ and selected shape parameters are given in Figure 1. It is a unimodal density function for various values of the shape parameters. It is easy to show that if $\alpha \to 0$ then (4) converges to gamma(2, λ) and if $\alpha \to \infty$ then (4) converges to exp(λ). Note that, when $\beta = 0$, then EWE(α, $\beta = 1$, λ) = WE(α, λ).

Interpretation 1: EWE distribution can be obtained the same way that Azzalini obtained the skew-normal distribution. Suppose X_1 and X_2 are two independent variables and $X_1 \sim \exp(\lambda)$, $X_2 \sim \mathrm{EE}(\lambda, \beta)$. For any $\alpha > 0$, consider a new random variable $X = X_1$ given that $\alpha X_1 > X_2$. It can be easily observed that the density function of X is (4).

Interpretation 2: EWE distribution can be obtained by the hidden truncation model proposed by Arnold and Beaver (2000). Suppose Z and Y are two dependent random variables with the joint density function

$$\mathrm{f}_{Z,Y}\left(z,y\right)=\frac{\lambda^{3}z}{\lambda+\beta}\left(1+\beta zy\right)\mathrm{e}^{-\lambda z\left(y+1\right)}, \quad z>0, y>0 \qquad (5)$$

It can be shown that the conditionally random variable $Z\,|Y \le \alpha$ has the EWE distribution.

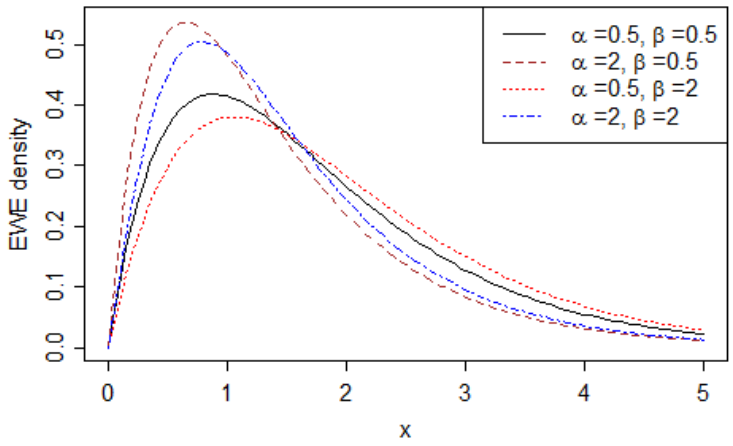


Figure 1. Plots of the EWE density function for fixed scale parameter $\lambda = 1$ and some selected shape parameters

Interpretation 3: Using the moment generating function (MGF) the stochastic representation of X can be easily obtained. Suppose U and V are two independent variables with distributions exp(λ) and EE($\lambda(1 + \alpha)$, $\alpha\beta$), respectively. Then it can be observed that if

$$X = U + V \tag{6}$$

then X has the density function (4).

Generation: All the above three interpretations can be used to generating random numbers from EWE distribution. Note that the simplest way to generate EWE random number is to use the stochastic representation (6).

Statistical and Reliability Properties

If $X \sim$ EWE(α, β, λ), then the MGF of X for any $t < \lambda$ is given by

$$\mathrm{M}_x(t) = \left(\frac{\lambda}{\lambda - t}\right)\left(\frac{\lambda(1+\alpha)}{\lambda(1+\alpha) - t}\right)^2\left(\frac{\lambda(1+\alpha) + \alpha\beta - t}{\lambda(1+\alpha) + \alpha\beta}\right) \tag{7}$$

By straightforward integration, the row moments of X about the origin are found to be

$$\mathrm{E}(X') = \frac{r!\left[(\lambda+\beta)(1+\alpha)^{r+2} - (\lambda+\beta)(1+\alpha) - \alpha\beta(r+1)\right]}{\lambda^{r}\alpha\left(\lambda(1+\alpha)+\beta\right)(1+\alpha)^{r}} \tag{8}$$

In particular, mean and $\mathrm{E}(X^2)$ are given, respectively, by

$$\mathrm{E}(X) = \frac{(\lambda+\beta)(1+\alpha)^{3} - (\lambda+\beta)(1+\alpha) - 2\alpha\beta}{\lambda\alpha\left(\lambda(1+\alpha)+\beta\right)(1+\alpha)} \tag{9}$$

$$\mathrm{E}(X^2) = \frac{2\left[(\lambda+\beta)(1+\alpha)^{4} - (\lambda+\beta)(1+\alpha) - 3\alpha\beta\right]}{\lambda^{2}\alpha\left(\lambda(1+\alpha)+\beta\right)(1+\alpha)^{2}} \tag{10}$$

The distribution function for the random variable X is given by

$$\begin{aligned} \mathrm{F}_X(x;\alpha,\beta,\lambda) = 1 - \mathrm{C}_0(\alpha,\beta,\lambda)\mathrm{e}^{-\lambda x} + \mathrm{C}_1(\alpha,\beta,\lambda)\mathrm{e}^{-\lambda(\alpha+1)x} \\ + \mathrm{C}_2(\alpha,\beta,\lambda)x\mathrm{e}^{-\lambda(\alpha+1)x} \end{aligned} \tag{11}$$

where

$$\mathrm{C}_0(\alpha,\beta,\lambda) = \frac{(1+\alpha)^{2}(\lambda+\beta)}{\alpha\left(\lambda(1+\alpha)+\alpha\beta\right)}, \quad \mathrm{C}_1(\alpha,\beta,\lambda) = \frac{(1+\alpha)(\lambda+\beta)+\alpha\beta}{\alpha\left(\lambda(1+\alpha)+\alpha\beta\right)},$$

$$\mathrm{C}_2(\alpha,\beta,\lambda) = \frac{\beta\lambda(1+\alpha)}{\left(\lambda(1+\alpha)+\alpha\beta\right)}$$

Also, the survival function and hazard rate function (HRF) of X can be placed in the following compact forms respectively:

$$\begin{aligned} \bar{\mathrm{F}}_X(x;\alpha,\beta,\lambda) = \mathrm{C}_0(\alpha,\beta,\lambda)\mathrm{e}^{-\lambda x} - \mathrm{C}_1(\alpha,\beta,\lambda)\mathrm{e}^{-\lambda(\alpha+1)x} \\ - \mathrm{C}_2(\alpha,\beta,\lambda)x\mathrm{e}^{-\lambda(\alpha+1)x} \end{aligned} \tag{12}$$

$$h_X(x)=\frac{\frac{(1+\alpha)^2\lambda}{\alpha(\lambda(1+\alpha)+\alpha\beta)}e^{-\lambda x}\left(\lambda+\beta-(\lambda+\beta+\alpha\beta\lambda x)^{-\alpha\lambda x}\right)}{C_0(\alpha,\beta,\lambda)e^{-\lambda x}-C_1(\alpha,\beta,\lambda)e^{-\lambda(\alpha+1)x}-C_2(\alpha,\beta,\lambda)xe^{-\lambda(\alpha+1)x}} \quad (13)$$

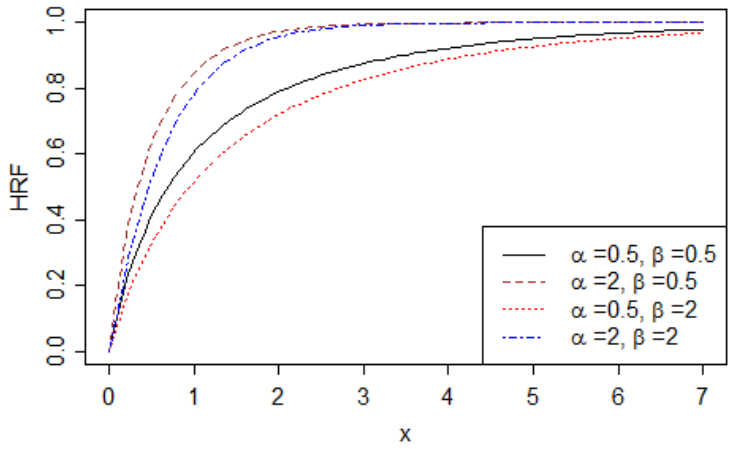


Figure 2. Plots of the EWE hazard rate function for fixed scale parameter $\lambda = 1$ and some selected shape parameters

In Figure 2, the HRF of the EWE distribution is plotted for selected values of the shape parameters and fixed scale parameter $\lambda = 1$. The HRF is an increasing function. The concept of an increasing failure rate is very attractive in an engineering context, where it has often been related to a mathematical representation of wear out (Marshall & Olkin, 2007).

One of the well-known properties of the life time distribution is mean residual life time. For the EWE distribution it can be written as

$$\begin{aligned} \mathrm{m}(t) &= \mathrm{E}(X-t \mid X>t) \\ &= \frac{1}{\lambda} \times \frac{\left[\begin{array}{c} \mathrm{C}_0(\alpha,\beta,\lambda)\mathrm{e}^{-\lambda t} - \mathrm{C}_3(\alpha,\beta,\lambda)\mathrm{e}^{-\lambda(\alpha+1)t} \\ -\dfrac{1}{1+\alpha}\mathrm{C}_2(\alpha,\beta,\lambda)t\mathrm{e}^{-\lambda(\alpha+1)t} \end{array}\right]}{\left[\begin{array}{c} \mathrm{C}_0(\alpha,\beta,\lambda)\mathrm{e}^{-\lambda t} - \mathrm{C}_1(\alpha,\beta,\lambda)e^{-\lambda(\alpha+1)t} \\ -\mathrm{C}_2(\alpha,\beta,\lambda)te^{-\lambda(\alpha+1)t} \end{array}\right]} \end{aligned} \tag{14}$$

where $\mathrm{C}_3(\alpha,\beta,\lambda) = \dfrac{(1+\alpha)(\lambda+\beta)+2\alpha\beta}{\alpha(\alpha+1)(\lambda(1+\alpha)+\alpha\beta)}$.

Maximum Likelihood Estimation

The MLEs will be derived for the unknown parameters of the EWE distribution from complete samples only.

Let $X_1,\ldots,X_n$ be a random sample from GWE(α, β, λ). The log-likelihood function based on the observed sample $\{x_1,\ldots,x_n\}$ is

$$\begin{aligned} \mathrm{l}(x_1,\ldots,x_n \mid \alpha,\beta,\lambda) &= 2n\ln(1+\alpha) + n\ln\lambda - n\ln\alpha - n\ln(\lambda(1+\alpha)+\alpha\beta) \\ &\quad -\lambda\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\ln\left(\lambda+\beta-(\lambda+\beta+\alpha\beta\lambda x_i)e^{-\alpha\lambda x_i}\right) \end{aligned} \tag{15}$$

To find the MLE estimates for the EWE model parameters, differentiate the log-likelihood function and equating the resulting expressions to 0 as follows:

$$\frac{\partial \mathrm{l}}{\partial\alpha} = \frac{2n}{1+\alpha} - \frac{n}{\alpha} - \frac{n(\lambda+\beta)}{\lambda(1+\alpha)+\alpha\beta} + \sum_{i=1}^{n}\frac{\alpha\beta\lambda^2x_i^2\mathrm{e}^{-\alpha\lambda x_i} - \beta\lambda x_i\mathrm{e}^{-\alpha\lambda x_i}}{\lambda+\beta-(\lambda+\beta+\alpha\beta\lambda x_i)\mathrm{e}^{-\alpha\lambda x_i}} = 0$$

$$\frac{\partial \mathrm{l}}{\partial\beta} = \frac{-n\alpha}{\lambda(1+\alpha)+\alpha\beta} + \sum_{i=1}^{n}\frac{1-(1+\alpha\lambda x_i)\mathrm{e}^{-\alpha\lambda x_i}}{\lambda+\beta-(\lambda+\beta+\alpha\beta\lambda x_i)\mathrm{e}^{-\alpha\lambda x_i}} = 0$$

$$\frac{\partial \mathrm{l}}{\partial\lambda} = \frac{n}{\lambda} - \frac{n(1+\alpha)}{\lambda(1+\alpha)+\alpha\beta} + \sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\frac{1-(1+\alpha\beta x_i)\mathrm{e}^{-\alpha\lambda x_i} + \alpha\lambda x_i(1+\alpha\beta x_i)\mathrm{e}^{-\alpha\lambda x_i}}{\lambda+\beta-(\lambda+\beta+\alpha\beta\lambda x_i)\mathrm{e}^{-\alpha\lambda x_i}} = 0$$

The MLEs of the unknown parameters cannot be obtained explicitly. They have to be obtained by solving some numerical methods, like the Newton-

Raphson method, Gauss-Newton method, or their variants. In this paper we use the *optim* function from the statistical software R (R Core Team, 2013) to estimate the unknown parameters.

Simulation

Some simulation results are presented to see how the maximum likelihood estimators behave for different sample sizes and for different parameter values. The sample sizes, namely $n = 20$, 40, 60, and 80 and two different sets of parameter values: Set 1: $\alpha = 0.5$, $\lambda = \beta = 1$, and Set 2: $\beta = 0.5$, $\alpha = \lambda = 1$. In each case, the maximum likelihood estimators of the unknown parameters are computed by maximizing the log-likelihood function (15). The average estimates and mean squared errors were computed over 1000 replications and the results are reported in Table 1. In all the cases the performances of the maximum likelihood estimates are quite satisfactory. As sample size increases the average estimates and the mean squared error decrease for all the parameters, as expected. It verifies the consistency properties of the MLEs.

Table 1. The average MLEs and the associated square root of the mean squared errors (within brackets)

	Set 1			Set 2		
n	***α***	***β***	***λ***	***α***	***β***	***λ***
20	0.6143	1.1257	1.1014	1.1316	0.6013	1.1286
	(0.0726)	(0.0793)	(0.0563)	(0.0811)	(0.0701)	(0.0599)
40	0.5825	1.1094	1.0614	1.1105	0.5784	1.0742
	(0.0592)	(0.0696)	(0.0352)	(0.0713)	(0.0501)	(0.0431)
60	0.5675	1.0835	1.0452	1.0922	0.0553	1.0562
	(0.0411)	(0.0536)	(0.0261)	(0.0658)	(0.0398)	(0.0371)
80	0.5595	1.0658	1.0352	1.0715	0.5462	1.0402
	(0.0388)	(0.0414)	(0.0201)	(0.0456)	(0.0321)	(0.0245)

Data Analysis

Two real data sets are considered to demonstrate the performance of the proposed distribution in practice. For each data set, the results of the fitted proposed model are compared with the WE, TWE, and EE models. To see which one of these models is more appropriate to fit the data set, the MLEs of unknown parameters and Akaike information criterion (AIC) were computed. The Kolmogorov-

Smirnov (K-S) distance between the empirical cumulative distribution function and the fitted distribution function was obtained in each case, as well as the associated p-value.

Data Set 1: Bjerkedal (1960) provided a data set consisting of survival times of 72 Guinea pigs injected with different amount of tubercle. We consider only the study in which animals in a single cage are under the same regimen. The data represents the survival times of Guinea pigs in days. The data are given below:

12 15 22 24 24 32 32 33 34 38 38 43 44 48 52 53 54 54 55 56 57 58 58 59 60 60 60 60 61 62 63 65 65 67 68 70 70 72 73 75 76 76 81 83 84 85 87 91 95 96 98 99 109 110 121 127 129 131 143 146 146 175 175 211 233 258 258 263 297 341 341 376

Table 2. The MLEs of parameters, AIC, and K-S statistics for the Guinea pigs data

Model	MLE of the parameters	AIC	K-S statistics	p-value
EE(β, λ)	10.1738, 0.0200	792.6086	0.1334	0.1544
WE(α, λ)	1.6312, 0.0138	791.1381	0.1153	0.2939
TWE(α_1, α_2, λ)	2.8013, 2.8013, 0.0142	789.0153	0.1132	0.3147
EWE(α, β, λ)	3.9035, 3.0313, 0.0141	**788.7657**	**0.1129**	**0.3174**

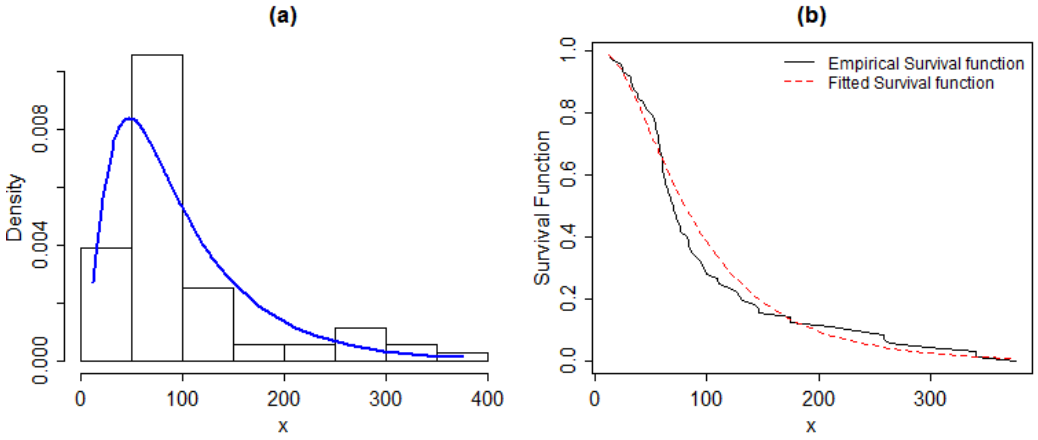


Figure 3. The fitted EWE distribution and the relative histogram for the Guinea pigs data (a); Empirical and fitted survival functions for the Guinea pigs data (b)

Table 3. The MLEs of parameters, AIC, and K-S statistics for the melanoma data

Model	MLE of the parameters	AIC	K-S statistics	*p*-value
EE(β, λ)	4.1321, 0.0019	913.2957	0.1551	0.1158
WE(α, λ)	1.6197 ,0.0010	912.5643	0.0767	0.8651
TWE(α_1, α_2, λ)	0.0136 0.2099 0.0022	913.3586	0.0696	0.9271
EWE(α, β, λ)	0.0375, 0.1435, 0.0021	**912.0859**	**0.0620**	**0.9710**

It is clear from Table 2 that, based on the AIC value and also based on the K-S statistic, the proposed EWE model provides a better fit than the WE, TWE and EE models for this specific data set. The relative histogram and the fitted EWE distribution are plotted in Figure 3. In order to assess if the model is appropriate, the plots of the fitted EWE survival function and empirical survival function are displayed in Figure 3.

Data Set 2: This data set relates to survival time for 57 patients in Denmark with malignant melanoma (Andersen, Borgan, Gill, & Keiding, 1993). The data are given below:

185 204 210 232 279 295 386 426 469 529 621 629 659 667 718 752 779 793
817 833 858 869 872 967 977 982 1041 1055 1062 1075 1156 1228 1252 1271
1312 1435 1506 1516 1548 1560 1584 1621 1667 1690 1726 1933 2061 2062
2103 2108 2256 2388 2467 2565 2782 3042 3338

The results are given in Table 3. The lowest values of the AIC and K-S test statistics are obtained for the EWE distribution. Based on these measures, the EWE is the best distribution among all those used here to fit the data set. In order to assess if the model is appropriate, the histogram of the data and the plot of the fitted EWE model are displayed in Figure 4.

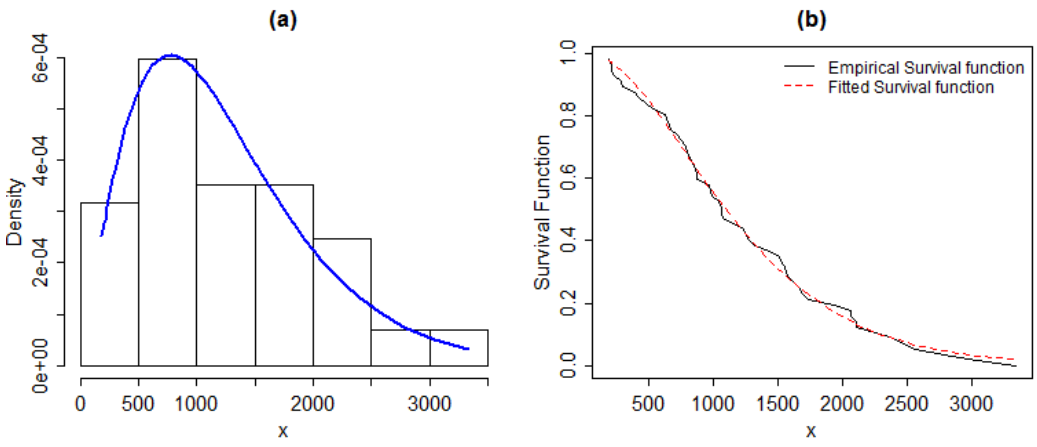


Figure 4. The fitted EWE distribution and the relative histogram for the melanoma data (a); Empirical and fitted survival functions for melanoma data (b)

Conclusion

A new class of weighted distributions based on the extended exponential distribution were introduced. The proposed model contains the WE model as its submodel. It is shown that the distribution function, hazard function, and moment generating function can be obtained in closed form. The MLEs can be computed using numerical algorithms. The failure rate function of proposed distributions is an increasing function. The flexibility of the proposed distribution and increased range of skewness was able to fit and capture features in two real data sets much better than the WE and other popular distributions.

References

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

Arnold, B. C., & Beaver, R. J. (2000). The skew-Cauchy distribution. Statistics & Probability Letters, 49(3), 285-290. doi: 10.1016/S0167-7152(00)00059-6

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171-178. Available from http://www.jstor.org/stable/4615982

Balakrishnan, N., & Ambagaspitiya, R. (1994). *On skew-Laplace distributions* (Technical report). Hamilton, Ontario, Canada: Department of Mathematics and Statistics, McMaster University.

Bjerkedal, T. (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli. *American Journal of Hygiene, 72*(1), 130-148.

Gómez, Y. M., Bolfarine, H., & Gómez, H. W. (2014). A new extension of the exponential distribution. *Revista Colombiana de Estadística, 37*(1), 25-34. doi: 10.15446/rce.v37n1.44355

Gupta, R. D., & Kundu, D. (2009). A new class of weighted exponential distributions. *Statistics: A Journal of Theoretical and Applied Statistics, 43*(6), 621-634. doi: 10.1080/02331880802605346

Marshall, A. W., & Olkin, I. (2007). *Life distributions: Structure of nonparametric, semiparametric, and parametric families*. New York: Springer.

Nadarajah, S. (2009). The skew logistic distribution. *AStA Advances in Statistical Analysis, 93*(2), 187-203. doi: 10.1007/s10182-009-0105-6

R Core Team. (2013). R: A language and environment for statistical computing [computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Shakhatreh, M. K. (2012). A two-parameter of weighted exponential distributions. *Statistics and Probability Letters, 82*(2), 252-261. doi: 10.1016/j.spl.2011.10.008

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 308-323.
doi: 10.22237/jmasm/1493597820

ISSN 1538 − 9472

Methodology For Constructing Perceptual Maps Incorporating Measuring Error In Sensory Acceptance Tests

Elisa Norberto Ferreira Santos
Fed. Univ. of Triângulo Mineiro
Uberaba, Brazil

Gilberto Rodrigues Liska
Federal University of Lavras
Lavras, Brazil

Marcelo Angelo Cirrillo
Federal University of Lavras
Lavras, Brazil

A new method is proposed based on construction of perceptual maps using techniques of correspondence analysis and interval algebra that allow specifying the measurement error expected in panel choices in the evaluation form described in unstructured 9-point hedonic scale.

Keywords: Interval algebra, correspondence analysis, panelist

Introduction

Sensory analysis is important in many domains: to improve the quality of products throughout the development process, to describe sensory properties of products, and to compare products to competitor's products (Latreille et al., 2006). Murray, Delahunty & Baxter (2001) treated the importance of descriptive sensory tests, noting that the sensory scientist requires an arsenal of sophisticated tools (Lawless & Heymann, 2010) to be applied to the detection (discrimination) and description of both the qualitative and quantitative sensory components of a consumer product by a trained panels of judges (see also Meilgaard, Civille & Carry, 1999). The qualitative aspects of a product include aroma, appearance, flavor, texture, aftertaste, and sound properties, and distinguish it from others. Sensory judges quantify these product aspects in order to facilitate description of the perceived product attributes.

There are several different methods of descriptive analysis: for instance, quantitative descriptive analysis (Stone & Sidel, 1993). Rossi (2001) suggested

Elisa Norberto Ferreira Santos is Faculty teaching statistics in the Department of Agronomic Engineering. Email at matematica.uab@iftm.edu.br. Email Gilberto Rodrigues Liska at gilbertoliska@hotmail.com. Email Marcelo Angelo Cirillo at macufla@dex.ufla.br.

repeatability and reproducibility measures defined by Mandel (1991). Others proposed more elaborate methodologies based on univariate or multivariate analysis with graphical and tabular representations of results.

Acceptance tests are generally applied to assess how much the consumer likes or dislikes a particular product (Prescott, 2009; Menezes et al., 2012). Different numerical scales are used for this purpose, especially the hedonic scale. Lim (2011), however, stated measurements of sensory or hedonic responses are inherent to effects relating to sensory and cognitive processes.

The stimulus-response model allows the interpretation that the first phase of sensory process, involving input of a stimulus, causes a sensory signal shown by feelings expressing quality and/or intensity. With regard to cognitive process, the initial phase is the decision that involves choice of scale, resulting in a more precise response to a specific sensory attribute, among other factors.

The relationship between sensory perceptions (sensory processing) and hedonic experience (cognitive process) is mentioned in the model as internal representation. Individual responses are certainly featured in a descriptive study summarized in numerical data. (Lim & Fujimaru, 2010). As to interference of the contextual effect in stimulus-response model, consider a situation where sensory perception comes from a trained panel with the ability to detect small differences between samples. Based on this panel's observations, and also considering the homogeneity of results obtained by a trained panel, results will certainly be more accurate than those of an untrained panel, which may show fatigue and unwillingness to perform all the tests, as well as heterogeneity in their skills and sensory perceptions. These are all important factors contributing to inaccurate responses.

Another factor that contributes to inaccuracy of answers is that responses from this range in practice are treated as continuous points. This suggests that parametric statistics such as analysis of variance may return incoherent results (Peryam & Pilgrim, 1957), because the assumptions are generally violated. See Gay & Mead (1992), Giovanni & Pangborn (1983), Lim, Wood & Green (2009), Lim & Fujimaru (2010), O'Mahony (1982), and Villanueva, Petenate and Silva (2000).

To find consumers who have similar liking patterns, clustering techniques have often been used (Yenket et al., 2011a; Liggett et al., 2008; Carlucci et al., 2009; Ares et al., 2010; Neely et al., 2010; Schmidt et al., 2010; Sinesio et al., 2010). Furthermore, to avoid the shortcomings inherent in the points system, new descriptive methodologies, such as the Quantitative Descriptive Analysis (QDA) have been developed (Stone & Sidel, 1993).

The advantages of QDA over other methods of evaluation are: (1) confidence in judgment of 10-12 trained panelists, instead of a few experts, (2) development of objective description closer to consumer language, and (3) consensual development of descriptive terminology, which implies higher concordance in judgments among panelists.

Amorim et al. (2010) indicated a good sensory panel should provide results that are accurate, discriminating, and precise. Thus, in a successful analysis, it is key to have a set of robust tools for monitoring individual assessor's performances as well as the performance of the panel as a whole. The success of using a sensory panel depends on its performance, i.e., its ability to identify small differences between products in certain attributes with statistical significance (Kermit & Lengard, 2005).

A good panel performance is achieved when each panelist discriminates between products (large product variability), repeats the assessments (small within-assessor variability) and agrees with all other panelists on the sensory sensation that is described by a particular attribute with certain strength (small between-assessor variability) (Derndorfer et al., 2005). Sample size estimation has been discussed (Gacula & Singh, 1984; Moskowitz, 1997; Lawless & Heymann, 2010; Gacula & Rutenbeck, 2006) over the last twenty years. It can be concluded that sample size calculation is generally an approximation because the formula contains elements based on assumptions such as the variance in the data and amount to be detected. Sensory scales vary in length; as a result, the variance and amount to be detected become a problem.

The sample or base size used in consumer acceptance tests has varied in practice, mostly based on experienced for a particular product. Thus, the proposed methodology is to construct perceptual maps with techniques of correspondence analysis (Blasius et al., 2009) that allow specification of the measurement error expected in relation to consumer/panelist choices in the evaluation form, described in an unstructured 9cm-point hedonic scale through interval algebra (Gioia & Lauro, 2005, 2006).

To illustrate this methodology, a case study is presented on sensory acceptance, considering different numbers of panelists in the evaluation of three genotypes of soybeans [Glycine max (L.) Merrill] called Black (MGBR07-7141), Brown (BRSMG-800A) and Yellow Soybeans (BRSMG-790A).

The statistical methodology proposed is applied to sensory acceptance tests, and has the advantages of quantitative descriptive analysis (QDA). The accuracy of the response interval is inferred by panelists, considering the expected measurement error in relation to consumer/panelist choices in the evaluation form

(described in unstructured hedonic terms). Usually, unstructured line scales are constructed, and a sample set is used to train panelists to reliably score the intensity of the chosen attributes.

Description of procedure for performing sensory tests applied to three soybean genotypes

Genotypes of soybeans [Glycine max (L.) Merrill] fit for human consumption in many seed coat colors came from the breeding program of the Embrapa/Epamig /Triângulo Foundation partnership, and sensory tests were performed at the Sensory Analysis Laboratory, Federal Institute IFTM-Triângulo Mineiro - Campus Uberaba, Brazil. The three genotypes were named according to the seed coat colors: Black (MGBR07-7141), Brown (BRSMG-800A), and Yellow Soybeans (BRSMG-790A).

Soybean genotypes were first soaked for 10 hours and then cooked with twice their volume of water. Cooking time was about 45 minutes in a pressure cooker, where each breed was cooked separately until they reached softness. Then the beans were cooled to approximately 25°C and served without spices. Acceptance test was conducted with 50 potential consumers of soybeans among students, teachers and administrative staff at IFTM, aged between 15-50 years, both genders.

The analysis was performed in individual white-lighted booths and samples were served in white plastic cups with a three-digit code. Six grains were served in each container and water was supplied to cleanse the palate between samples. Grains were presented in monadic sequential scheme (one at a time) in unstructured 9cm-hedonic scale from 1 (dislike extremely) to 9 (like extremely) to assess appearance, texture, and overall acceptance.

Incorporation of fundamentals of interval algebra in correspondence analysis and construction of perceptual maps

Based on the panelist scores obtained, the concepts of interval algebra were incorporated into sensory analysis considering each score and giving a measurement error $\xi = \pm 0.2$ cm and $\xi = \pm 1.0$ cm, which was determined by a priori knowledge of the researchers.

In agreement with the statistical methodology and given the unstructured 9-point hedonic scale, imposition of measurement error ξ to be made by the

panelists in marking the acceptance form was made by considering two conjectures. First, the panelists showed some similar sensory abilities, i.e., there is a slight error in marking, arbitrarily set at $\xi = \pm 0.2$ cm, to be considered in measuring results. Second, the panelists show some heterogeneous sensory abilities, i.e., there was an error of considerable extent, arbitrarily set at $\xi = \pm 1.0$ cm, to be considered in measuring results.

Importantly, the accuracy of each measurement depended on the skills of panelists. No matter how careful the measurement and how precise the scoring in the evaluation form, there was always an uncertainty due to panel heterogeneity. However, as scoring uncertainty is considered when using interval algebra for constructing perceptual maps, both inaccuracy and accuracy of scores become predictable. Therefore, it is consistent to use a smaller sample size in acceptance testing. Thus, considering 50 panelists for each sensory attribute, each interval observation was represented by $\left[\underline{f}_{ij}; \overline{f}_{ij} \right]$ for the i^{th} taster ($i = 1, ..., I = 50$) and j^{th} cultivate ($j = 1, ..., J = 3$), the lower limit $\underline{f}_{ij}$ being calculated by the score $ij - \xi$ and the upper limit $\overline{f}_{ij}$ represented by the score $ij + \xi$.

Thus, interval sensory data were organized in a contingency table of interval frequency for constructing perceptual maps (Table 1) in a way similar to correspondence analysis (Guedes et al., 1999).

Table 1. Contingency table of interval frequency used for constructing perceptual maps

Panelist n(*i*)	Genotypes of Soybeans			Total
	Black (MGBR07-7141)	Yellow (BRSMG-790A)	Brown (BRSMG-800A)	
n_1	$\left[\underline{f}_{11}; \overline{f}_{11} \right]$	$\left[\underline{f}_{12}; \overline{f}_{12} \right]$	$\left[\underline{f}_{13}; \overline{f}_{13} \right]$	$\left[\sum_{j=1}^{J} \underline{f}_{1j}; \sum_{j=1}^{J} \overline{f}_{1j} \right]$
n_2	$\left[\underline{f}_{21}; \overline{f}_{21} \right]$	$\left[\underline{f}_{22}; \overline{f}_{22} \right]$	$\left[\underline{f}_{23}; \overline{f}_{23} \right]$	⋮
⋮	⋮	⋮	⋮	⋮
n_I	$\left[\underline{f}_{I1}; \overline{f}_{I1} \right]$	$\left[\underline{f}_{I2}; \overline{f}_{I2} \right]$	$\left[\underline{f}_{I3}; \overline{f}_{I3} \right]$	⋮
Total	$\left[\sum_{i=1}^{I} \underline{f}_{i1}; \sum_{i=1}^{I} \overline{f}_{i1} \right]$	...	...	$\left[\sum_{i=1}^{I} \sum_{j=1}^{J} \underline{f}_{ij}; \sum_{i=1}^{I} \sum_{j=1}^{J} \overline{f}_{ij} \right]$

Following the structure of the interval data shown in Table 1, we obtained the correlation matrix considering interval data (1).

$$[Q]=\begin{bmatrix} \left[\underline{q}_{11};\overline{q}_{11}\right] & \left[\underline{q}_{12};\overline{q}_{12}\right] & \left[\underline{q}_{13};\overline{q}_{13}\right] \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \left[\underline{q}_{I1};\overline{q}_{I1}\right] & \left[\underline{q}_{I2};\overline{q}_{I2}\right] & \left[\underline{q}_{IJ};\overline{q}_{IJ}\right] \end{bmatrix} \tag{1}$$

where each element was calculated by the expression (2) following specific mathematical operations for interval division (Gioia & Lauro, 2005).

$$\left[\underline{q}_{ij};\overline{q}_{ij}\right]=\frac{\left[\underline{f}_{ij};\overline{f}_{ij}\right]}{\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\underline{f}_{ij};\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{f}_{ij}\right]} \text{ for } i=1,...,I;\ j=1,...,J \tag{2}$$

After obtaining the correlation matrix considering data interval, use the chi-square correction which resulted in the matrix **[*D*]**, each element being obtained by (3).

$$d_{ij}=\frac{\left[\underline{q}_{ij};\overline{q}_{ij}\right]-\left[\underline{q}_{i.};\overline{q}_{i.}\right]\left[\underline{q}_{.j};\overline{q}_{.j}\right]}{\sqrt{\left[\underline{q}_{i.};\overline{q}_{i.}\right]\left[\underline{q}_{.j};\overline{q}_{.j}\right]}} \tag{3}$$

where marginal probabilities were respectively defined for lines and columns of the correlation matrix considering data interval, according to expressions (4) and (5).

$$\left[\underline{q}_{i.};\overline{q}_{i.}\right]=\begin{bmatrix}\left[\sum_{j=1}^{J}\underline{q}_{1j};\sum_{j=1}^{J}\overline{q}_{1j}\right]\\ \left[\sum_{j=1}^{J}\underline{q}_{2j};\sum_{j=1}^{J}\overline{q}_{2j}\right]\\ \vdots\\ \left[\sum_{j=1}^{J}\underline{q}_{Ij};\sum_{j=1}^{J}\overline{q}_{Ij}\right]\end{bmatrix} \tag{4}$$

$$\left[\underline{q}_{.j};\overline{q}_{.j}\right]=\left[\left[\sum_{i=1}^{I}\underline{q}_{i1};\sum_{i=1}^{I}\overline{q}_{i1}\right]\ \left[\sum_{i=1}^{I}\underline{q}_{i2};\sum_{i=1}^{I}\overline{q}_{i2}\right]\ \cdots\ \left[\sum_{i=1}^{I}\underline{q}_{iJ};\sum_{i=1}^{I}\overline{q}_{iJ}\right]\right] \tag{5}$$

Interval mathematical operations used for calculating probabilities were performed as described by Gioia & Lauro (2005). Thus, regarding the correlation matrix considering data interval **[*D*]**, whose dimension is *I* lines by *J* columns, corrected by the chi-squared distance, covariance matrices associated with profiles 'line' and 'column' keeping interval data were respectively determined by (6) and (7).

$$\left[\Sigma_{L}\right]=\left[D\right]^{T}\left[D\right] \tag{6}$$

$$\left[\Sigma_{C}\right]=\left[D\right]\left[D\right]^{T} \tag{7}$$

The normalization procedures used for profiles 'line' and 'column' were performed with singular value decomposition (Gioia & Lauro, 2006; Deif & Rohn, 1994; Seif, Hashem & Deif, 1992) considering the matrices $\mathbf{[\Sigma_L]}$ and $\mathbf{[\Sigma_C]}$ whose dimension is *I* lines by *J* columns. The position of each profile 'line' in relation to profiles 'column' were obtained in (8) and (9).

$$\left[L\right]=\left[D_{L}\right]^{-\frac{1}{2}}\left[U\right] \tag{8}$$

where $\mathbf{[D_L]^{-½}}$ is the square root of the diagonal matrix of the marginal probabilities 'line' of **[*Q*]** and **[*U*]** is the matrix of normalized eigenvectors of $\mathbf{[\Sigma_L]}$. Similarly, the position of each profile 'column' in relation to profiles 'line' was determined by

$$[C] = [D_C]^{-\frac{1}{2}}[V] \quad (9)$$

where **[*V*]** is the matrix of eigenvectors normalized of $[\Sigma_C]$, and $[D_C]^{-½}$ is the square root of the diagonal matrix of marginal probabilities 'column' of **[*Q*]**.

Based on the interval matrices **[*L*]** and **[*C*]** the coordinates related to profiles 'line' were been given by $[\tilde{L}] = [D_L]^{-1}[Q]^T[C]$ and the coordinates related to profiles 'column' were obtained by $[\tilde{C}] = [D_C]^{-1}[Q]^T[L]$.

A total inertia of the cloud of points is illustrated in Figure 1.

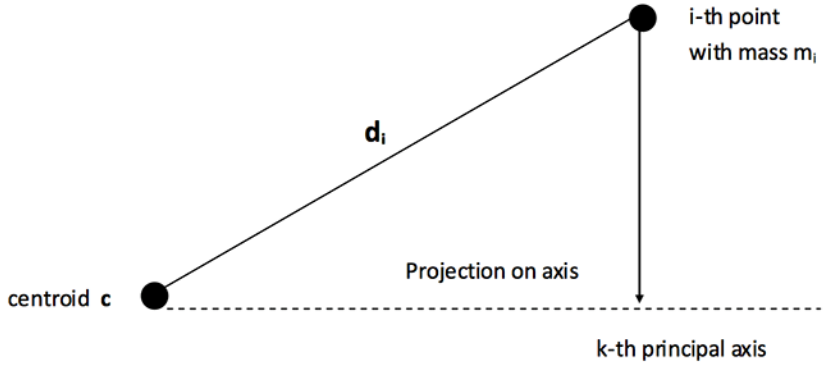


Figure 1. Inertia of decomposition in correspondence analysis

The coordinates obtained enabled the construction of interval perceptual maps, using a routine in R (R Core Team, 2013), and similar to technique preference maps as follows: coordinate values, variance explained on the first two components, consumer space, descriptive space, descriptive attributes that promote liking as recommended Yenket, et al. (2011b).

Results

Considering acceptance data in interval scale in relation to the attribute appearance, the results compiled in Figure 2 correspond to perceptual maps constructed respectively to $\xi = \pm 0.2$ cm (A) and $\xi = \pm 1.0$ cm (B). Percentage of sample variation explained for axes F1 and F2 is shown in Table 2.

Table 2. Decomposition of sample variability for the attribute appearance

	Axis	Inertia	Proportion	Cumulative (%)
(A) ξ = ± 0.2 cm	F1	[1.6918; 2.2516]	[0.8420; 0.8629]	[84.20; 86.29]
	F2	[0.2687; 0.4225]	[0.1370; 0.1579]	[97.90; 102.8]
	Total	[1.9605; 2.6741]		
(B) ξ = ± 1 cm	F1	[1.9584; 4.0786]	[0.5950; 1.742]	[59.50; 174.2]
	F2	[1.3326; 2.3408]	[0.3646; 0.4049]	[95.96; 214.69]
	Total	[3.2910; 6.4194]		

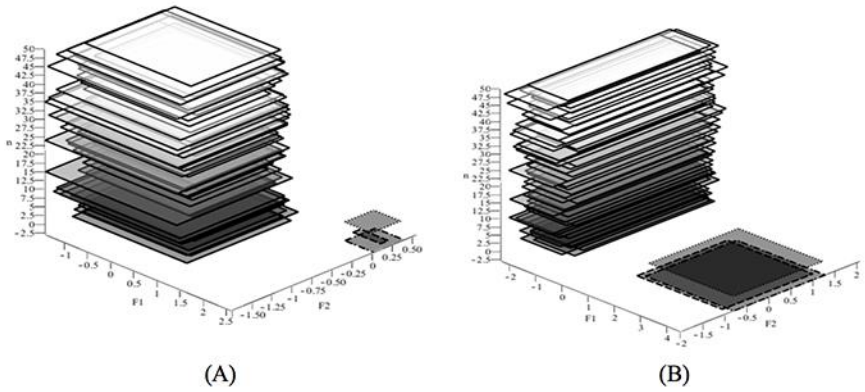


Figure 2. Perceptual map using interval scale for the attribute 'appearance'. Grayscale shows the 50 panelists, dotted line displays cultivar MGBR07-7141 (Black Soybeans), dash line for cultivar BRSMG-790A (Yellow Soybeans), and dashed-dotted line for cultivar BRSMG-800A (Brown Soybeans).

Results in Figure 2(A) indicated when considering a small measurement error ξ = ± 0.2 cm there is statistical evidence to state that the panel responses were homogeneous with respect to the attribute appearance, however, there was no evidence of preference for any particular soybean cultivar. Nevertheless, by increasing the measurement error to ξ = ± 1.0 cm, results in Figure 2(B) showed panel scores with a certain degree of similar homogeneity and no preference to cultivate, since a simple inspection of the rectangles indicated they had similar areas.

Given the two differential conjectures by different margins of error to be considered in response marking, and also keeping in mind the statement of Cohen (1990) related to beliefs and opinions of consumers about a product, such results

would most likely help companies develop packaging, labels, and advertising campaigns to inform consumers about characteristics and properties of products in order to raise consumer expectations and encourage purchase. Thus, constructing perceptual maps via interval scaling definitely minimizes uncertainties regarding product acceptability as far as publicity is concerned.

Perceptual maps for evaluation of the attribute overall acceptance are described in Figure 3, while percentage of sample variation explained for axes F1 and F2 is shown in Table 3.

Table 3. Decomposition of sample variability for the attribute overall acceptance

	Axis	Inertia	Proportion	Cumulative (%)
(A) ξ = ± 0.2 cm	F1	[1.4175; 2.8151]	[0.7120; 0.8189]	[71.20; 81.89]
	F2	[0.3133; 1.1386]	[0.1810; 0.2879]	[89.3; 110.68]
	Total	[1.7308; 3.9537]		
(B) ξ = ± 1 cm	F1	[1.0985; 2.6511]	[0.4706; 0.5616]	[47.06; 56.16]
	F2	[0.8572; 2.9814]	[0.4383; 0.5293]	[90.89; 109.09]
	Total	[1.9557; 5.6325]		

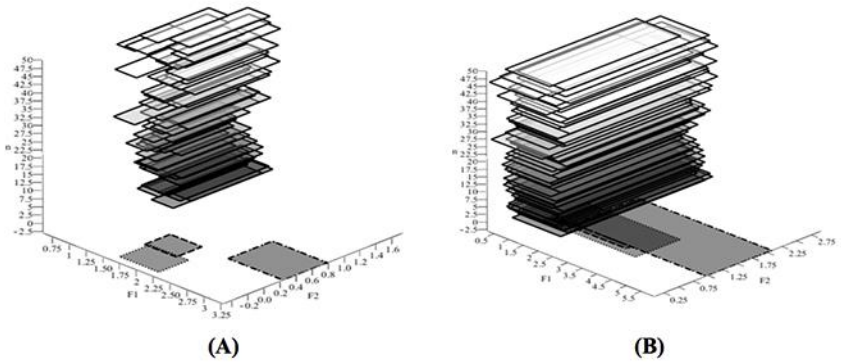


Figure 3. Perceptual map using interval scale for the attribute 'overall acceptance'. Grayscale shows the 50 panelists, dotted line displays cultivar MGBR07-7141 (Black Soybeans), dash line for cultivar BRSMG-790A (Yellow Soybeans), and dash-dotted line for cultivar BRSMG-800A (Brown Soybeans).

Considering the situation of a small and essential error in response marking represented by $\xi = \pm 0.2$ cm (Figure 3(A)), a greater heterogeneity is seen between panelists. However, cultivar preference is inconclusive with regard to the attribute overall acceptance, as rectangle areas look similar. When considering the conjecture in which scale variability is greater, results in Figure 3(B) indicated homogeneous panel scores, although showing no specific preference for any particular soybean cultivar, as the rectangles do not overlap. Yenket et al. (2011a) mentioned this may be based on the frequency of a particular product being most or least liked by individual consumers and is not based on mean liking scores for a group of consumers.

Using perceptual maps reinforces the hypothesis that incorporating measurement error in data analysis is recommended provided there is a priori knowledge of the critical values for the margin of error. However, not all errors have to be measured. Behrens & Silva (2004) stated that the score given to the attribute 'overall acceptance' is merely determined by a simple inspection. Also, the response is related to the panelist attitude influenced by individual learning and experience on the object of our study: soybean genotypes, degree of individual acceptance/preference, and motivational component associated with action tendency. Perceptual maps for evaluation of the attribute 'texture' are shown in Figure 4, while percentage of sample variation explained for axes F1 and F2 is shown in Table 4.

Table 4. Decomposition of sample variability for the attribute texture

	Axis	Inertia	Proportion	Cumulative (%)
(A) $\xi = \pm 0.2$ cm	F1	[1.3216; 1.6500]	[0.7698; 0.9402]	[76.98; 94.02]
	F2	[0.3950; 0.1048]	[0.0597; 0.2301]	[82.95; 117.03]
	Total	[1.7166; 1.7548]		
(B) $\xi = \pm 1$ cm	F1	[1.1440; 4.4134]	[0.6067; 0.6319]	[60.67; 63.19]
	F2	[0.7414; 2.5701]	[0.3680; 0.3932]	[97.47; 102.51]
	Total	[1.8854; 6.9835]		

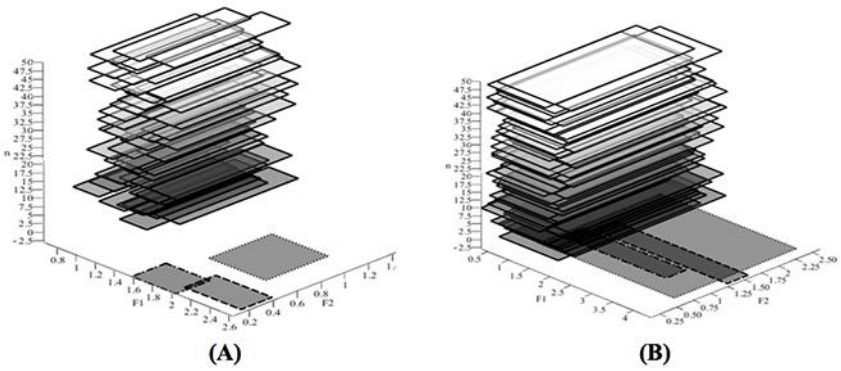


Figure 4. Perceptual map using interval scale for the attribute 'texture'. Grayscale shows the 50 panelists, dotted line displays cultivar MGBR07-7141 (Black Soybeans), dash line for cultivar BRSMG-790A (Yellow Soybeans), and dashed-dotted line for cultivar BRSMG-800A (Brown Soybeans).

Results plotted in Figure 4(A) showed that scores for the attribute texture were very different, considering that the panelists could have made a mistake of $\xi = \pm 0.2$ cm when marking answers. Thus, there is no evidence of preference for any particular soybean cultivar, as rectangles do not overlap. In the situation with the greatest measurement error, arbitrarily set at $\xi = \pm 1.0$ cm, the results in Figure 4(B) indicated more homogeneous scores, which showed evidence of similarity among the genotypes BRSMG-790A (Yellow Soybeans) and BRSMG-800A (Brown Soybeans). This was evidenced by overlapping in most areas of cultivar-specific rectangles. Score differentiation regarding the genotype MGBR07-7141 (Black Soybeans) could possibly be influenced by physiological aspects, as seed coat is very important for regulating water absorption.

McDonald Jr. et al. (1988) stated that water intake affects a few morphological characteristics of seed coats that may influence water penetration time. Thus, it is reasonable to assume that physicochemical properties of genotypes with different seed coat colors are differentiated. This fact could possibly imply a genotype appearance more or less pleasing to the panelists, either in appearance or texture, so that responses of sensory evaluations presumably could be influenced by stimulation effect (Lim, 2011). Such effect is impossible to detect by incorporating measurement error, as the contextual interference effect suggested by Lim, Wood, and Green (2009) was recognized as a source of error and bias in evaluation testing.

Conclusion

Different scale variability in the case study showed that using interval algebra in correspondence analysis applied to descriptive tests provided additional information on the accuracy of panelist responses. Concerning the selection of soybean genotypes, incorporating measurement error in data analysis allowed for identification of groups with similar genotypes due to subjective analysis of profile location and overlapping in the quadrants.

Acknowledgments

The authors thank CNPq and the Fapemig project for financial assistance.

References

Amorim, I. S., Ferreira, E. B., Lima, R. R., and Pereira, R. G. F. A. (2010). Monte Carlo based test for inferring about the unidimensionality of a Brazilian coffee sensory panel. *Food Quality and Preference, 21*, 319–323. doi: 10.1016/j.foodqual.2009.08.018

Ares, G., Barreiro, C., Deliza, R., Gimémes, A. & Gámbaro, A. (2010). Application of a check-all-that-apply question to the development of chocolate milk desserts. *Journal of Sensory Studies, 25*, 67–86. doi: 10.1111/j.1745-459X.2010.00290.x

Behrens, J. H. & Silva, M. A. A. P. (2004). Consumer attitude towards soybean and related products. *Food Science and Technology, 24*(3), 431-439. doi: 10.1590/S0101-20612004000300023.

Blasius, J., Greenacre, M., Groenen, P. J. F. & Velden, M. V. (2009). Special issue on correspondence analysis and related methods. *Computational Statistics & Data Analysis, 53*, 3103-3106. doi: 10.1016/j.csda.2008.11.010

Carlucci, A., Monteleone, E., Braghieri, A. & Napolitano, F. (2009). Mapping the effect of information about animal welfare on consumer liking and willingness to pay for yogurt. *Journal of Sensory Studies, 24*, 712–730. doi: 10.1111/j.1745-459X.2009.00235.x

Cohen, J. C. (1990). Applications of qualitative research for sensory analysis and product development. *Food Technology, 44*(11), 164-166.

Deif, A. S. & Rohn, J. (1994). On the invariance of the sign pattern of matrix eigenvectors under perturbation. *Linear Algebra and its Applications, 196*, 63-70. doi: 10.1016/0024-3795(94)90315-8

Derndorfer, E., Baierl, A., Nimmervoll, E. & Sinkovits, E. (2005). A panel performance procedure implemented in R. *Journal of Sensory Studies, 20*(3), 217–227. doi: 10.1111/j.1745-459X.2005.00021.x

Gacula, M. C. & Singh, J. (1984). *Statistical Methods in Food and Consumer Research*. Orlando: Academic Press FL.

Gacula, M. Jr. & Rutenbeck, S. (2006). Sample size in consumer test and descriptive analysis. *Journal of Sensory Studies, 21*(2), 129–145. doi: 10.1111/j.1745-459X.2006.00055.x

Gay, C. & Mead, R. (1992). A statistical appraisal of the problem of sensory measurement. *Journal of Sensory Studies, 7*(3), 205-228. doi: 10.1111/j.1745-459X.1992.tb00533.x

Gioia, F. & Lauro, C. N. (2005). Basic statistical method for interval data. *Statistica Applicata, 17*(1), 75-104.

Gioia, F. & Lauro, C. N (2006). Principal component analysis on interval data. *Computational Statistics, 21*, 343–363. doi: 10.1007/s00180-006-0267-6

Giovanni, M. E. & Pangborn, R. M. (1983). Measurement of taste intensity and degree of liking of beverages by graphic scales and magnitude estimation. *Journal of Food Science, 48*(4), 1175-1182. doi: 10.1111/j.1365-2621.1983.tb09186.x

Guedes, A. T., Ivanqui, I. L., Martins, A. B. T. & Cochia, E. B. R. (1999). Seleção de variáveis categóricas utilizando análise de correspondência e análise procustes. *Acta Scientiarum, 21*(1), 861-868. doi: 10.4025/actascitechnol.v21i0.3084

Kermit, M. & Lengard, V. (2005). Assessing the performance of a sensory panel – Panelist monitoring and tracking. *Journal of Chemometrics, 19*(3), 154-161. doi: 10.1002/cem.918

Latreille, J. et al. (2006). Measurement of the reliability of sensory panel performances. *Food Quality and Preference, 17*, 369-375. doi 10.1016/j.foodqual.2005.04.010

Lawless, H. T. & Heymann, H. (2010). *Sensory evaluation of food: principles and practices*. New York: Chapman and Hall. doi: 10.1007/978-1-4419-6488-5

Liggett, R. E., Drake, M. A. & Delwiche, J. F. (2008). Impact of flavor attributes on consumer liking of Swiss cheese. *Journal of Dairy Science, 91*(2), 466–476. doi: 10.3168/jds.2007-0527

Lim, J., Wood, A. & Green, B. G. (2009). Derivation and evaluation of a labeled hedonic scale. *Chemical Senses, 34*(9), 739-751. doi: 10.1093/chemse/bjp054

Lim, J. & Fujimaru, T. (2010). Evaluation of the labeled hedonic scale under different experimental conditions. *Food Quality and Preference, 21*(5), 521-530. doi: 10.1016/j.foodqual.2010.02.001

Lim, J. (2011). Hedonic scaling: a review of methods and theory. *Food Quality and Preference, 22*(8), 733-747. doi: 10.1016/j.foodqual.2011.05.008

Mandel, J. (1991). The validation of measurement through interlaboratory studies. *Chemometrics and Intelligent Laboratory Systems, 11*(2), 109-119. doi: 10.1016/0169-7439(91)80058-X

McDonald Jr., M. B., Vertucci, C. W. & Roos, E. C. (1988). Soybean seed imbibition: water absorption by seed parts. *Crop Science, 28*(6), 993-997. doi: 10.2135/cropsci1988.0011183X002800060026x

Meilgaard, M.C., Civille, G.V., and Carr, B.T. (1999). *Sensory evaluation techniques* (3rd Ed.). Boca Raton, FL: CRC Press. doi: 10.1201/9781439832271

Menezes, C. C. , Borges, S. , Carneiro, J. D. , Cirillo, M. A. & Oliveira, L. F. (2012). Optimization of Guava (Psidium guajava, L) preserves using the acceptance test, response surface methodology and preference mapping. *Boletim do Centro de Pesquisa e Processamento de Alimentos, 30*(1), 1-10. doi: 10.5380/cep.v30i1.28654

Moskowitz, H. R. (1997). Base size in product testing: A psychophysical viewpoint and analysis. *Food Quality and Preference, 8*(4), 247–255. doi: 10.1016/S0950-3293(97)00003-7

Murray, J. M., Delahunty, C. M. & Baxter, I. A. (2001). Descriptive Sensory analysis: past, present, and future. *Food Research International, 34*(6), 461-471. doi: 10.1016/S0963-9969(01)00070-9

O'Mahony, M. (1982). Some assumptions and difficulties with common statistics for sensory analysis. *Food Technology, 36*, 75-82.

Neely, E. A., Lee, Y. & Lee, S-Y. (2010). Cross-cultural comparison of acceptance of soy-based extruded snack foods by U.S. and Indian consumers. *Journal of Sensory Studies, 25*, 87–108. doi: 10.1111/j.1745-459X.2010.00276.x

Peryam, D. R. & Pilgrim, F. J. (1957). Hedonic scale method of measuring food preference. *Food Technology, 11*, 9-14.

Prescott, J. (2009). Rating a new hedonic scale: a commentary on "derivation and evaluation of a labeled hedonic scale" by Lim, Wood and Green. *Chemical Senses, 34*(9), 735-737. doi: 10.1093/chemse/bjp072

R Core Team (2013). *R: A language and environment for statistical computing* [computer program]. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/

Rossi, F. (2001). Assessing sensory panelist performance using repeatability and reproducibility measures. *Food Quality and Preference, 12*, 467-479. doi: 10.1016/S0950-3293(01)00038-6

Seif, N. P., Hashem, S. & Deif, A. S. (1992). Bounding the eigenvectors for symmetric interval matrices. *Journal of Applied Mathematics and Mechanics, 72*(3), 233-236. doi: 10.1002/zamm.19920720313

Schmidt, T. B., Schilling, M. W., Behrends, J. M., Battula, V., Jackson, V., Sekhon, R. K. & Lawrence, T. E. (2010). Use of cluster analysis and preference mapping to evaluate consumer acceptability of choice and select bovine M. longissimus lumborum steaks cooked to various end-point temperatures. *Meat Science, 84*(1), 46–53. doi: 10.1016/j.meatsci.2009.08.016

Sinesio, F., Cammareri, M., Moneta, E., Navez, B., Peparaio, M., Causse, M. & Grandillo, S. (2010). Sensory quality of fresh French and Dutch market tomatoes: A preference mapping study with Italian consumers. *Journal of Food Science, 75*(1), 55–67. doi: 10.1111/j.1750-3841.2009.01424.x

Stone, H. S. & Sidel, J. L. (1993). *Sensory Evaluation Practices*. (2nd ed). San Diego, CA: Academic Press.

Villanueva, N. D. M., Petenate, A. J. & Silva, M.A.A.P. (2000). Performance of three affective methods and diagnosis of the anova model. *Food Quality and Preference, 11*(5), 363-370. doi: 10.1016/S0950-3293(00)00006-9

Yenket, R., Chambers, E. I. V. & Johnson, D. E. (2011a). Statistical package clustering may note be best for grouping consumers to understand their most liked products. *Journal of Sensory Studies, 26*(3), 209-225. doi: 10.1111/j.1745-459X.2011.00337.x

Yenket, R., Chambers, E. I. V. & Adhikari, K. (2011b). A Comparison of seven preference mapping techniques using four software programs. *Journal of Sensory Studies, 26*(2), 135–150. doi: 10.1111/j.1745-459X.2011.00330.x

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 324-349.
doi: 10.22237/jmasm/1493597880

ISSN 1538 − 9472

Confidence Intervals for the Scaled Half-Logistic Distribution under Progressive Type-II Censoring

Kiran G. Potdar
Ajara Mahavidyalaya
Ajara, India

D. T. Shirke
Shivaji University
Kolhapur, India

Confidence interval construction for the scale parameter of the half-logistic distribution is considered using four different methods. The first two are based on the asymptotic distribution of the maximum likelihood estimator (MLE) and log-transformed MLE. The last two are based on pivotal quantity and generalized pivotal quantity, respectively. The MLE for the scale parameter is obtained using the expectation-maximization (EM) algorithm. Performances are compared with the confidence intervals proposed by Balakrishnan and Asgharzadeh via coverage probabilities, length, and coverage-to-length ratio. Simulation results support the efficacy of the proposed approach.

Keywords: Progressively Type-II censoring, EM algorithm, MLE, pivotal quantity, confidence interval, generalized confidence interval, coverage probability, coverage to length ratio, half-logistic distribution

Introduction

In many life testing situations, an experiment has to be terminated before completion. Because of the various limitations of time and money, testing of life may need to be stopped for some of the units. In day-to-day experiments, incomplete information about the failure times is available, or some of the units must be removed before completion of the experiment. A plan is necessary for removal of the units before the termination of an experiment to save time and cost, which is called the censored data.

Type-I censoring depends on time, where the time is fixed for the termination of experiment. Suppose an observer continues an experiment up to time T; lifetimes of units will be known exactly only if these are less than T.

Dr. Potdar is an Assistant Professor in the Department of Statistics. Email them at: potdarkiran.stat@gmail.com. Dr. Shirke is a Professor in the Department of Statistics. Email them at: dts_stats@unishivaji.ac.in.

Failure times of units which have not failed by the time T are not observed. Suppose n units are being tested, but the decision is made to terminate the experiment at time T. In this experiment, lifetimes will be known exactly only for those units that fail before time T. In Type-I censoring, the number of exact lifetimes observed is random.

A Type-II censoring scheme is often used in life testing experiments where the number of units that can be observed before the termination of the experiment is fixed. In this scheme, only a pre-planned number m out of n units ($m < n$) are observed. In the case of Type-II censoring, the number of exact lifetimes observed is fixed, but the time required for the termination of the experiment is unknown. In conventional Type-I and Type-II censoring, units are removed from the experiment at the terminal stage, while in a progressive censoring scheme, units are removed at different stages. Progressive censoring schemes can be applied in both Type-I and Type-II censoring schemes. More details about various censoring schemes are available in Lawless (1982).

In an $(R_1, R_2,\ldots, R_m)$ progressive type-II censoring scheme, the number m and $R_1, R_2,\ldots, R_m$ are fixed before the start of the experiment and $\sum_{i=1}^{m} R_i = n-m$. At the first failure, R_1 units are randomly removed from the remaining $n-1$ units. At the second failure, R_2 units are randomly removed from the remaining $n-2-R_1$ units, etc. At the m^{th} failure, all the remaining R_m units are removed. Here, we observe failure times of m units and the remaining $n-m$ units are removed at different stages of the experiment. In a conventional Type-II censoring scheme, $R_m = n - m$ and the rest of the R_i are zero.

Consider the problem of interval estimation for the scale parameter of a half-logistic distribution under a progressive Type-II censoring scheme. Progressive Type-II censoring schemes for various lifetime distributions was discussed by Cohen (1963), who introduced progressive Type-II censoring schemes. Mann (1969, 1971), Balakrishnan, Kannan, Lin, and Ng (2003), Balakrishnan, Kannan, Lin, and Wu (2004), Ng (2005), and Ng, Kundu, and Balakrishnan (2006) discussed inference for different lifetime distributions under progressive Type-II censoring schemes. Balakrishnan and Aggarwala (2000) is an excellent reference on progressive censoring. Balakrishnan (2007) studied various distributions and inferential methods for the progressively censored data. Lin and Balakrishnan (2011) discussed the consistency and the asymptotic normality of Maximum Likelihood Estimators (MLEs) based on the progressive Type-II censored samples. Potdar and Shirke (2013, 2014) studied inference for the scale parameter of the half logistic and Rayleigh distribution of k-unit parallel systems

based on progressively Type-II censored data. Ghitany, Alqallaf, and Balakrishnan (2014) discussed estimation of the parameters of Gompertz distributions based on progressively Type-II censored samples. Sultan, Alsadat, and Kundu (2014) studied estimation for the inverse Weibull parameters under progressive Type-II censoring.

As far as the half-logistic distribution is concerned, Balakrishnan and Puthenpura (1986) discussed the best linear unbiased estimation of location and scale parameters. Balakrishnan and Wong (1991) computed the approximate Maximum Likelihood Estimator (AMLE) for the location and scale parameters of the half-logistic distribution. Balakrishnan and Chan (1992) studied estimation for the scale parameter of the half-logistic distribution. Kim and Han (2010) used importance sampling methods to obtain a Bayes estimator for the scale parameter of the half-logistic distribution under progressively Type-II censored samples. Jang, Park, and Kim (2011) studied estimation of the scale parameter of the half-logistic distribution with a multiply Type-II censored sample. Rastogi and Tripathi (2014) studied estimation of parameter and reliability for the exponentiated half-logistic distribution.

The likelihood equation of a half-logistic distribution with scale parameter does not have a closed form solution to obtain MLE. In most of the reported work, an AMLE of the scale parameter is obtained. Following this approach, Balakrishnan and Asgharzadeh (2005) and Wang (2009) reported inference for the scale parameter of a half-logistic distribution based on progressive Type-II censored samples.

Balakrishnan and Asgharzadeh (2005) showed that, if the relative sample fraction is small, then the coverage probability of the confidence interval (CI) based on asymptotic normality of the MLE is unsatisfactory. Wang (2009) paid more attention to length of CI and gave a shorter length CI. Dempster, Laird, and Rubin (1977) introduced the expectation-maximization (EM) algorithm to obtain the MLE for the incomplete data. McLachlan and Krishnan (1997) gave more details about the EM algorithm. Here, the MLE is computed using the EM algorithm, and the focus is on both the coverage probability and length of CI.

Assume that n units having half-logistic lifetime distribution are put on test and failure times of $\sum_{i=1}^{m} R_i = n - m$ units are censored. Lifetimes of these censored units are unknown. Consider the censored data as missing data and use the EM algorithm to compute the MLE. As indicated in Potdar and Shirke (2014), the EM algorithm gives improved inferential results.

Model and Estimation of the Scale Parameter

Suppose progressively Type-II censored data are obtained from the scaled half-logistic distribution with probability density function

$$\mathrm{f}\left(x;\lambda\right)=\frac{1}{\lambda}\frac{2\mathrm{e}^{-x/\lambda}}{\left(1+\mathrm{e}^{-x/\lambda}\right)^{2}},\quad x\geq 0,\lambda>0 \tag{1}$$

and cumulative distribution function

$$\mathrm{F}\left(x;\lambda\right)=\left[\frac{1-\mathrm{e}^{-x/\lambda}}{1+\mathrm{e}^{-x/\lambda}}\right],\quad x\geq 0,\lambda>0$$

Suppose n units are under test and lifetimes of m units are observed under progressive Type-II censoring. Suppose (R_1, R_2,…, R_m), a progressive censoring scheme, is used. The observed lifetimes $x_{(1)}$, $x_{(2)}$,…, $x_{(m)}$ are the progressively Type-II censored sample. The likelihood function for the observed data is given by (Balakrishnan & Aggarwala, 2000)

$$\mathrm{L}\left(\lambda\right)=C\prod_{i=1}^{m}\mathrm{f}\left(x_{(i)};\lambda\right)\left[1-\mathrm{F}\left(x_{(i)};\lambda\right)\right]^{R_i} \tag{2}$$

where

$$C=n\prod_{j=1}^{m-1}\left(n-j-\sum_{i=1}^{j}R_i\right)$$

Maximum Likelihood Estimation

Suppose $\mathbf{z}_1$, $\mathbf{z}_2$,…,$\mathbf{z}_m$ are the censored data. Note $\mathbf{z}_i$ is a vector with R_i element corresponding to R_i removed units after the i^{th} failure is observed (i = 1, 2,…., m). The censored data $\mathbf{Z}$ = ($\mathbf{z}_1$, $\mathbf{z}_2$,…, $\mathbf{z}_m$) can be considered to be the missing data and $\mathbf{X}$ = ($x_{(1)}$, $x_{(2)}$,…, $x_{(m)}$) the observed data. $\mathbf{W}$ = ($\mathbf{X}$, $\mathbf{Z}$) is the complete data set to be used for drawing inference for the scale parameter. The complete log-likelihood function can be written as

$$L_c = -n\log(\lambda) + \sum_{i=1}^{m}\log\left(\frac{2e^{-x_i/\lambda}}{\left(1+e^{-x_i/\lambda}\right)^2}\right) + \sum_{i=1}^{m}\sum_{j=1}^{R_i}\log\left(\frac{2e^{-z_{ij}/\lambda}}{\left(1+e^{-z_{ij}/\lambda}\right)^2}\middle| z_{ij} > x_i\right) \quad (3)$$

By differentiating L_c with respect to λ,

$$\frac{dL_c}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i\left(1-e^{-x_i/\lambda}\right)}{1+e^{-x_i/\lambda}} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\left(\frac{z_{ij}\left(1-e^{-z_{ij}/\lambda}\right)}{1+e^{-z_{ij}/\lambda}}\middle| z_{ij} > x_i\right)$$

The EM algorithm suggested by Dempster et al. (1977) was used to compute the MLE. For the E step in the EM algorithm, the expectation of Z_{ij} was taken. Hence, the above equation becomes

$$\frac{dL_c}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i\left(1-e^{-x_i/\lambda}\right)}{1+e^{-x_i/\lambda}} + \frac{1}{\lambda^2}\sum_{i=1}^{m}R_i\,a(x_i,\lambda) \quad (4)$$

where

$$a(x_i,\lambda) = E\left(\frac{z_{ij}\left(1-e^{-z_{ij}/\lambda}\right)}{1+e^{-z_{ij}/\lambda}}\middle| z_{ij} > x_i\right) = \int_{x_i}^{\infty}\frac{z\left(1-e^{-z_{ij}/\lambda}\right)}{1+e^{-z_{ij}/\lambda}}\frac{f(z)}{1-F(x_i)}dz$$

$$= \frac{1+e^{-x_i/\lambda}}{\lambda e^{-x_i/\lambda}}\int_{x_i}^{\infty}\frac{ze^{-z/\lambda}\left(1-e^{-z/\lambda}\right)}{\left(1+e^{-z/\lambda}\right)^3}dz$$

Solving equation (4) is the M step.

The Newton-Raphson method was used to solve equation (4) by taking the least square estimate as an initial value. Ng (2005) discussed estimation of model parameters of modified Weibull distributions based on progressively Type-II censored data, where the empirical distribution function is computed as

$$\hat{F}\left(x_{(i)}\right) = 1 - \prod_{j=1}^{i}\left(1-\hat{p}_j\right)$$

with

$$\hat{p}_j = \frac{1}{n - p_j^*},\ p_j^* = \sum_{k=2}^{j} R_{k-1} - j + 1, \quad j = 1, 2, \ldots, m$$

The estimate of the parameters can be obtained by the least squares fit of simple linear regression

$$y_i = \beta x_{(i)}$$

with $\beta = -1/\lambda$,

$$y_i = \ln\left[\frac{1 - \frac{\hat{\mathrm{F}}\left(x_{(i-1)}\right) + \hat{\mathrm{F}}\left(x_{(i)}\right)}{2}}{1 + \frac{\hat{\mathrm{F}}\left(x_{(i-1)}\right) + \hat{\mathrm{F}}\left(x_{(i)}\right)}{2}}\right], \quad i = 1, 2, \ldots, m$$

$$\hat{\mathrm{F}}\left(x_{(0)}\right) = 0$$

The least square estimate of λ is given by

$$\hat{\lambda}_0 = -\frac{\sum_{i=1}^{m} x_{(i)}^2}{\sum_{i=1}^{m} x_{(i)} y_i} \tag{5}$$

While obtaining the MLE $\hat{\lambda}_n$ of the scale parameter λ, the above approach was adopted, where $\hat{\lambda}_0$ was taken as an initial value of λ in the Newton-Raphson method. It will be shown that the MLE $\hat{\lambda}_n$ exits and is unique. From equation (2),

$$\mathrm{L}(\lambda) = C \prod_{i=1}^{m} \frac{2^{R_i+1}}{\lambda} \frac{\mathrm{e}^{-(R_i+1)x_i/\lambda}}{\left(1 - \mathrm{e}^{-x_i/\lambda}\right)^{R_i+2}}$$

where C is defined as above.

$$\frac{d\log \mathrm{L}}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\left(R_i+1\right)x_i - \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{\left(R_i+2\right)x_i \mathrm{e}^{-x_i/\lambda}}{1+x_i \mathrm{e}^{-x_i/\lambda}} = 0 \tag{6}$$

Define

$$\mathrm{g}\left(\lambda\right) = \frac{-\lambda^2 d\log \mathrm{L}}{d\lambda} = \frac{d\log \mathrm{L}}{d\lambda} = n\lambda - \sum_{i=1}^{m}\left(R_i+1\right)x_i + \sum_{i=1}^{m}\frac{\left(R_i+2\right)x_i \mathrm{e}^{-x_i/\lambda}}{1+x_i \mathrm{e}^{-x_i/\lambda}} = 0$$

Note

$$\lim_{\lambda\to 0} g\left(\lambda\right) < 0, \lim_{\lambda\to\infty} g\left(\lambda\right) > 0, \text{ and } \mathrm{g}'\left(\lambda\right) > 0$$

Therefore, the MLE, a solution to g(λ) = 0, exists and is unique.

Fisher Information

We compute observed Fisher information using the idea of the missing information principle of Louis (1982). Thus, observed information = complete information – missing information. Write this as

$$\mathrm{I}_x\left(\lambda\right) = \mathrm{I}_w\left(\lambda\right) - \mathrm{I}_{w|x}\left(\lambda\right) \tag{7}$$

In the following, we obtain complete and missing information given by

$$\mathrm{I}_w\left(\lambda\right) = -\mathrm{E}\left[\frac{d^2 \mathrm{L}}{d\lambda^2}\right]$$

where, L is the log-likelihood function of the complete data. By differentiating L with respect to λ twice

$$\frac{d^2 \mathrm{L}}{d\lambda^2} = \frac{n}{\lambda^2} - \frac{2}{\lambda^4}\sum_{i=1}^{n}\frac{x_i^2 \mathrm{e}^{-x_i/\lambda}}{\left(1+\mathrm{e}^{-x_i/\lambda}\right)^2} - \frac{2}{\lambda^3}\sum_{i=1}^{n}\frac{x_i\left(1-\mathrm{e}^{-x_i/\lambda}\right)}{1+x_i \mathrm{e}^{-x_i/\lambda}}$$

The complete information is given by

$$\mathrm{I}_w(\lambda) = -\frac{n}{\lambda^2} + \frac{2}{\lambda^4}\sum_{i=1}^{n}\mathrm{E}\left[\frac{X_i^2 \mathrm{e}^{-X_i/\lambda}}{\left(1+\mathrm{e}^{-X_i/\lambda}\right)^2}\right] + \frac{2}{\lambda^3}\sum_{i=1}^{n}E\left[\frac{X_i\left(1-\mathrm{e}^{-X_i/\lambda}\right)}{1+x_i\mathrm{e}^{-X_i/\lambda}}\right] \qquad (8)$$

Missing information is given by

$$\mathrm{I}_{W|X}(\lambda) = \sum_{i=1}^{m} R_i \, \mathrm{I}_{W|X}^{(i)}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i}\mathrm{E}_{Z|X}\left[\frac{d^2 \log\left(\mathrm{f}\left(Z_{ij} \mid X_i, \lambda\right)\right)}{d\lambda^2}\right]$$

Consider

$$\mathrm{f}_{z|X}\left(Z_{ij} \mid X_i, \lambda\right) = \frac{\mathrm{f}\left(z_{ij};\lambda\right)}{1-\mathrm{F}\left(x_i;\lambda\right)} = \frac{\frac{1}{\lambda}\frac{2\mathrm{e}^{-z_{ij}/\lambda}}{\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)^2}}{1-\left[\frac{1-\mathrm{e}^{-x_i/\lambda}}{1+\mathrm{e}^{-x_i/\lambda}}\right]}$$

Therefore,

$$\log \mathrm{f} = -\log\lambda + \log\left[\frac{2e^{-z_{ij}/\lambda}}{\left(1+e^{-z_{ij}/\lambda}\right)^2}\right] - \log\left[1-\left(\frac{1-e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right)\right]$$

$$\frac{d\log \mathrm{f}}{d\lambda} = -\frac{1}{\lambda} + \frac{z_{ij}\left(1-\mathrm{e}^{-z_{ij}/\lambda}\right)}{\lambda^2\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)} - \frac{x_i}{\lambda^2\left(1+\mathrm{e}^{-x_i/\lambda}\right)}$$

and

$$\frac{d^2\log \mathrm{f}}{d\lambda^2} = \frac{1}{\lambda^2} - \frac{2z_{ij}^2\mathrm{e}^{-z_{ij}/\lambda}}{\lambda^4\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)^2} - \frac{2z_{ij}\left(1-\mathrm{e}^{-z_{ij}/\lambda}\right)}{\lambda^3\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)} + \frac{2x_i^2\mathrm{e}^{-x_i/\lambda}}{\lambda^4\left(1+\mathrm{e}^{-x_i/\lambda}\right)^2} + \frac{2x_i}{\lambda^3\left(1+\mathrm{e}^{-x_i/\lambda}\right)}$$

Hence

$$\begin{aligned} \mathrm{I}_{W|X}(\lambda) &= \sum_{i=1}^{m} R_i \, \mathrm{I}_{W|X}^{(i)}(\lambda) \\ &= -\frac{n-m}{\lambda^2} + \frac{2}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\mathrm{E}\left[\frac{z_{ij}^2 \mathrm{e}^{-z_{ij}/\lambda}}{\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)^2}\right] + \frac{2}{\lambda^3}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\mathrm{E}\left[\frac{z_{ij}\left(1-\mathrm{e}^{-z_{ij}/\lambda}\right)}{1+\mathrm{e}^{-z_{ij}/\lambda}}\right] \quad (9) \\ &\quad -\frac{1}{\lambda^4}\sum_{i=1}^{m}\frac{R_i x_i^2 \mathrm{e}^{-x_i/\lambda}}{\left(1+\mathrm{e}^{-x_i/\lambda}\right)^2} - \frac{2}{\lambda^3}\sum_{i=1}^{m}\frac{R_i x_i}{\left(1+\mathrm{e}^{-x_i/\lambda}\right)} \end{aligned}$$

Confidence Intervals Based on MLE and log-Transformed MLE

Confidence Interval Based on MLE

Let $\hat{\lambda}_n$ be the MLE of λ and

$$\hat{\sigma}^2\left(\hat{\lambda}_n\right) = \frac{1}{\mathrm{I}\left(\hat{\lambda}_n\right)}$$

be the estimated asymptotic variance of $\hat{\lambda}_n$. Therefore, a 100(1 – α)% asymptotic CI for λ based on asymptotic normality of $\hat{\lambda}_n$ is given by

$$\left(\hat{\lambda}_n - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}, \hat{\lambda}_n + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}\right) \quad (10)$$

where $\tau_{\alpha/2}$ is the upper 100(α/2)$^{\text{th}}$ percentile of the standard normal distribution.

Confidence Interval Based on log-Transformed MLE

Meeker and Escobar (1998) reported the asymptotic CI for λ based on $\log\left(\hat{\lambda}_n\right)$. An approximate 100(1 – α)% CI for log(λ) is

$$\left(\log\left(\hat{\lambda}_n\right) - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right)}, \log\left(\hat{\lambda}_n\right) + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right)}\right)$$

where $\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right)$ is the estimated asymptotic variance of $\log\left(\hat{\lambda}_n\right)$, which is approximated by

$$\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right) \approx \frac{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}{\hat{\lambda}_n^2}$$

Hence, an approximate $100(1-\alpha)\%$ CI for λ is

$$\left(\hat{\lambda}_n e^{\left(-\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}}{\hat{\lambda}_n}\right)}, \hat{\lambda}_n e^{\left(\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}}{\hat{\lambda}_n}\right)}\right) \tag{11}$$

Confidence Interval Based on Pivotal and Generalized Pivotal Quantity

Consider two exact CIs based on the pivotal quantities. To define these CIs, show that the distribution of $V = \hat{\lambda}/\lambda$ is free from λ, where $\hat{\lambda}$ is the MLE of λ, based on the complete data. In the following lemma, it is proved that V is a pivot, following Gulati and Mi (2006):

Lemma 1: The distribution of V is free from λ.

Proof: Consider the probability density function of the half-logistic distribution with scale parameter λ:

$$f\left(x,\lambda\right) = \frac{2e^{-x/\lambda}}{\lambda\left(1+e^{-x/\lambda}\right)^2}, \quad x \geq 0, \lambda > 0$$

Then the log-likelihood function becomes

$$L = -n\log\left(\lambda\right) + n\log\left(2\right) - \frac{1}{\lambda}\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n}\log\left(1+e^{-x_i/\lambda}\right)$$

$dL/d\lambda = 0$ gives the following equation:

$$\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} \frac{x_i e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}} = n\lambda$$

The solution of the above equation is the MLE of λ (say $\hat{\lambda}$). Hence

$$\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} \frac{x_i e^{-x_i/\hat{\lambda}}}{1+e^{-x_i/\hat{\lambda}}} = n\hat{\lambda}$$

$$\sum_{i=1}^{n} \frac{x_i}{\lambda} - 2\sum_{i=1}^{n} \frac{x_i}{\lambda} \frac{e^{\frac{-x_i\lambda}{\lambda\hat{\lambda}}}}{1+e^{\frac{-x_i\lambda}{\lambda\hat{\lambda}}}} = \frac{n\hat{\lambda}}{\lambda}$$

Let $\xi = \lambda/\hat{\lambda}$ and $Y_i = X_i/\lambda$. Then

$$\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} y_i \frac{e^{-y_i\xi}}{1+e^{-y_i\xi}} = n\xi^{-1}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{2}{n}\sum_{i=1}^{n} y_i \frac{e^{-y_i\xi}}{1+e^{-y_i\xi}} - \xi^{-1} = 0$$

Note that $Y_1, Y_2, \ldots, Y_n$ is a random sample from the half-logistic distribution with parameter $\lambda = 1$. Therefore, the distribution of $\xi = \lambda/\hat{\lambda}$ is independent of λ. Hence the proof.

Lemma 2: The distribution of V under progressive Type-II censored data from the half-logistic distribution with scale parameter λ is free from λ.

Proof: This is similar to Lemma 1 and hence is omitted.

This property of the MLE will be used to derive the confidence interval based on pivot and generalized pivot quantity methods.

Remark: V is also a pivot for k-unit parallel and k-unit series systems.

Confidence Interval Based on Pivotal Quantity

From Lemma 2, the distribution of V is free from λ. Define a and b such that

$$\mathrm{P}\left(a < V < b\right) = 1 - \alpha$$

Therefore we obtain the following as a CI for λ:

$$\left(\frac{\hat{\lambda}}{b}, \frac{\hat{\lambda}}{a}\right) \tag{12}$$

The constants a and b are obtained using Monte Carlo simulation by using the following algorithm:

***Algorithm to Obtain Percentiles of* V**

1. Input α, N, m, and progressive Type-II censoring scheme (R_1, R_2,…, R_m).
2. Generate a progressive Type-II censored random sample of size m using censoring scheme (R_1, R_2,…, R_m) from the half-logistic distribution with parameter $\lambda = 1$.
3. Obtain a MLE of λ (say $\hat{\lambda}$) using the EM algorithm.
4. Repeat steps 2 and 3 N times so as to get $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_N$.
5. Arrange the $\hat{\lambda}_i$ in an increasing order. Denote them by $\hat{\lambda}_{(1)}, \hat{\lambda}_{(2)}, \ldots, \hat{\lambda}_{(N)}$.
6. Compute $a = \hat{\lambda}_{([(\alpha/2)N])}$ and $b = \hat{\lambda}_{([(1-\alpha/2)N])}$.

Confidence Interval Based on Generalized Pivotal Quantity

The concept of a generalized confidence interval (GCI) is introduced by Weerahandi (1993). Let x denote the observed value of X. To construct a GCI for λ, first define a generalized pivotal quantity (GPQ), $\mathrm{T}(X; x, \lambda)$, which is a function of the random variable X, its observed value x, and the parameter λ. A quantity $\mathrm{T}(X; x, \lambda)$ is required to satisfy the following two conditions:

i) For a fixed x, the probability distribution of $\mathrm{T}(X; x, \lambda)$ is free of unknown parameters.
ii) The observed value of $\mathrm{T}(X; x, \lambda)$, namely $\mathrm{T}(x; x, \lambda)$, is simply λ.

Let T_α be the $100\alpha^{\text{th}}$ percentile of T. Then T_α becomes the $100(1-\alpha)\%$ lower bound for λ. Therefore a $100(1-\alpha)\%$ two-sided GCI for parameter λ is given by

$$\left(\mathrm{T}_{\alpha/2}, \mathrm{T}_{1-\alpha/2}\right) \tag{13}$$

Define the GPQ as

$$\mathrm{T}\left(X; x, \lambda\right) = \frac{\hat{\lambda}_0}{\hat{\lambda}/\lambda}$$

where $\hat{\lambda}_0$ is the MLE obtained using observed data. Note:

i) The distribution of T(X; x, λ) is free from λ, which follows from Lemma 2, and
ii) T(x; x, λ) = λ, since for the observed data, $\hat{\lambda} = \hat{\lambda}_0$.

A GCI based on T(X; x, λ) is obtained by using following algorithm:

Algorithm to Obtain CI for λ using GPQ

1. Input α, N, m, and progressive Type-II censoring scheme (R_1, R_2,…, R_m).
2. Generate a progressive Type-II censored random sample of size m from the half-logistic distribution with an unknown parameter λ.
3. Based on the data in step 2, obtain a MLE of λ (say $\hat{\lambda}_0$) using the EM algorithm.
4. Generate a progressive Type-II censored random sample of size m from the half-logistic distribution with parameter $\lambda = 1$.
5. Obtain a MLE of λ (say $\hat{\lambda}_i$) using the EM algorithm for step 4 data.
6. Compute $T_i = \hat{\lambda}_0 / \hat{\lambda}_i$.
7. Repeat steps 4 to 6 N times, so as to get T_1, T_2,…,T_N.
8. Arrange the T_i in an increasing order. Denote them by $T_{(1)}$, $T_{(2)}$,…, $T_{(N)}$.
9. Compute a 100(1 – α)% CI for λ as $\left(T_{(\lceil(\alpha/2)N\rceil)}, T_{(\lceil(1-\alpha/2)N\rceil)}\right)$.

Simulation Study

The CIs given in (10) to (13) will now be compared with the CIs given by Balakrishnan and Asgharzadeh (2005) and Wang (2009). A simulation study was

carried out to study the performance of each of the CIs. Asymptotic CIs based on MLE, log-transformed MLE, and GPQ are compared through length and confidence level. Balakrishnan and Sandhu (1995) presented an algorithm for sample generation from progressively Type-II censored schemes. This algorithm was used to generate samples from a half-logistic distribution. Consider the 34 different progressively Type-II censored schemes compiled in Table 1.

Algorithm

1. Generate i.i.d. observations (W_1, W_2,…,W_m) from U(0, 1).
2. For censoring scheme (R_1, R_2,…, R_m),

$$E_i = \frac{1}{\left(i + R_m + R_{m-1} + \ldots + R_{m-i+1}\right)}$$

   for i = 1, 2,…, m.
3. Set $V_i = W_i^{E_i}$ for i = 1, 2,…,m.
4. Set $U_i = 1 - (V_m \cdot V_{m-1} \cdot \ldots \cdot V_{m-i+1})$ for i = 1, 2,…,m. Then (U_1, U_2,…, U_m) is the uniform (0, 1) progressively Type-II censored sample.
5. For given values of the parameter λ, set

$$x_{(i)} = -\lambda \log\left[\frac{1-U_i}{1+U_i}\right]$$

   for i = 1, 2,…, m.

Then ($x_{(1)}$, $x_{(2)}$,…, $x_{(m)}$) is the required progressively Type-II censored sample from the half-logistic distribution. In Table 1, censoring scheme (a, b, c, d) stands for $R_1 = a$, $R_2 = b$, $R_3 = c$, and $R_4 = d$. A similar meaning holds for schemes described through completely specified vector, while scheme (10, 9*0) means $R_1 = 10$ and remaining nine R_i are zero, i.e. $R_2 = R_3 = R_4 = \ldots = R_{10} = 0$. A simulation was carried out with $\lambda = 1$. For each particular progressive censoring scheme, 5,000 sets of observations are generated. The CIs based on asymptotic normal distributions of the MLE and log-transformed MLE are derived.

Table 1. Censoring schemes

Scheme No.	*n*	*m*	*m*/*n*	Scheme
[1]	10	4	0.2500	(0, 0, 0, 6)
[2]	10	4	0.2500	(6, 0, 0, 0)
[3]	10	5	0.5000	(0, 0, 0, 0, 5)
[4]	10	5	0.5000	(5, 0, 0, 0, 0)
[5]	15	4	0.2667	(0, 0, 0, 11)
[6]	15	4	0.2667	(11, 0, 0, 0)
[7]	15	5	0.3333	(0, 0, 0, 0, 10)
[8]	15	5	0.3333	(10, 0, 0, 0, 0)
[9]	15	5	0.3333	(0, 10, 0, 0, 0)
[10]	15	5	0.3333	(0, 0, 10, 0, 0)
[11]	15	5	0.3333	(2, 2, 2, 2, 2)
[12]	15	5	0.3333	(4, 4, 2, 0, 0)
[13]	20	5	0.2500	(0, 0, 0, 0, 15)
[14]	20	5	0.2500	(15, 0, 0, 0, 0)
[15]	20	5	0.2500	(5, 5, 5, 0, 0)
[16]	20	5	0.2500	(3, 3, 3, 3, 3)
[17]	20	5	0.2500	(0, 15, 0, 0, 0)
[18]	20	5	0.2500	(5, 10, 0, 0, 0)
[19]	20	10	0.5000	(9*0, 10)
[20]	20	10	0.5000	(10, 9*0)
[21]	25	5	0.2000	(0, 0, 0, 0, 20)
[22]	25	5	0.2000	(20, 0, 0, 0, 0)
[23]	25	10	0.4000	(9*0, 15)
[24]	25	10	0.4000	(15, 9*0)
[25]	25	15	0.6000	(14*0, 10)
[26]	25	15	0.6000	(10, 14*0)
[27]	50	20	0.4000	(19*0, 30)
[28]	50	20	0.4000	(30, 19*0)
[29]	50	25	0.5000	(24*0, 25)
[30]	50	25	0.5000	(25, 24*0)
[31]	100	20	0.2000	(19*0, 80)
[32]	100	20	0.2000	(80, 19*0)
[33]	100	50	0.5000	(49*0, 50)
[34]	100	50	0.5000	(50, 49*0)

Table 2. Simulated coverage probabilities for confidence intervals

	C_1		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[1]	0.8100	0.8396	0.8108	0.8470	0.8710	0.9176	0.8944	0.9458	0.8992	0.9474
[2]	0.8300	0.8640	0.8338	0.8676	0.8804	0.9282	0.9072	0.9514	0.8986	0.9464
[3]	0.8288	0.8638	0.8330	0.8684	0.8768	0.9256	0.8968	0.9462	0.9025	0.9503
[4]	0.8290	0.8688	0.8382	0.8768	0.8814	0.9286	0.9014	0.9528	0.9036	0.9494
[5]	0.8204	0.8508	0.8160	0.8500	0.8786	0.9204	0.8978	0.9476	0.9016	0.9518
[6]	0.8350	0.8650	0.8364	0.8706	0.8830	0.9306	0.8978	0.9528	0.8948	0.9468
[7]	0.8194	0.8582	0.8278	0.8640	0.8736	0.9230	0.8998	0.9522	0.9058	0.9548
[8]	0.8360	0.8686	0.8418	0.8778	0.8834	0.9284	0.9006	0.9528	0.8998	0.9482
[9]	0.8370	0.8684	0.8398	0.8724	0.8794	0.9240	0.9050	0.9526	0.8986	0.9498
[10]	0.8354	0.8656	0.8364	0.8666	0.8780	0.9306	0.8946	0.9456	0.8978	0.9506
[11]	0.8262	0.8596	0.8308	0.8684	0.8822	0.9274	0.9022	0.9494	0.9050	0.9518
[12]	0.8354	0.8650	0.8408	0.8798	0.8896	0.9336	0.9014	0.9514	0.8934	0.9486
[13]	0.8318	0.8626	0.8418	0.8750	0.8842	0.9348	0.9002	0.9504	0.8966	0.9520
[14]	0.8474	0.8806	0.8474	0.8834	0.8866	0.9342	0.8960	0.9474	0.8974	0.9462
[15]	0.8368	0.8740	0.8388	0.8716	0.8752	0.9250	0.8974	0.9528	0.9008	0.9482
[16]	0.8308	0.8632	0.8312	0.8664	0.8816	0.9260	0.9048	0.9532	0.8950	0.9496
[17]	0.8432	0.8724	0.8492	0.8818	0.8870	0.9296	0.9004	0.9504	0.9000	0.9464
[18]	0.8318	0.8690	0.8390	0.8756	0.8788	0.9260	0.8944	0.9488	0.8998	0.9500
[19]	0.8592	0.8954	0.8790	0.9122	0.8902	0.9416	0.8960	0.9510	0.8950	0.9458
[20]	0.8680	0.9068	0.8706	0.9098	0.8864	0.9358	0.9002	0.9528	0.8958	0.9418
[21]	0.8196	0.8544	0.8280	0.8606	0.8764	0.9284	0.8990	0.9496	0.8976	0.9492
[22]	0.8372	0.8720	0.8400	0.8712	0.8764	0.9304	0.8972	0.9542	0.8970	0.9504
[23]	0.8640	0.9072	0.8636	0.8994	0.8858	0.9364	0.8976	0.9490	0.8980	0.9454
[24]	0.8774	0.9128	0.8780	0.9132	0.8964	0.9434	0.8904	0.9466	0.9010	0.9512
[25]	0.8714	0.9160	0.8770	0.9158	0.8948	0.9432	0.8926	0.9448	0.9006	0.9466
[26]	0.8822	0.9210	0.8848	0.9242	0.8996	0.9504	0.9008	0.9492	0.8938	0.9468
[27]	0.8844	0.9246	0.8790	0.9212	0.8914	0.9388	0.9002	0.9502	0.8970	0.9472
[28]	0.8852	0.9302	0.8880	0.9292	0.8952	0.9470	0.9084	0.9532	0.8948	0.9496

Table 2, continued.

	C_1		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[29]	0.8904	0.9276	0.8950	0.9360	0.9022	0.9494	0.9024	0.9466	0.8948	0.9504
[30]	0.8896	0.9348	0.8918	0.9374	0.8982	0.9484	0.9044	0.9530	0.8978	0.9478
[31]	0.8920	0.9324	0.8856	0.9248	0.8962	0.9460	0.9008	0.9526	0.8968	0.9486
[32]	0.8864	0.9306	0.8876	0.9336	0.8972	0.9478	0.9062	0.9534	0.8958	0.9478
[33]	0.8930	0.9374	0.8938	0.9408	0.8998	0.9454	0.8958	0.9446	0.9046	0.9530
[34]	0.8924	0.9416	0.9010	0.9452	0.9026	0.9522	0.8948	0.9448	0.9070	0.9544

Table 3. The expected lengths of confidence intervals

	C_1		C_2		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[1]	2.0913	2.7742	2.0330	2.7028	1.3723	1.6352	1.4919	1.8397	2.0003	2.6406	2.0432	2.7096
[2]	2.0150	2.6663	1.9223	2.5345	1.3790	1.6432	1.4943	1.8403	1.9281	2.5360	1.9254	2.5328
[3]	1.6829	2.2413	1.6495	2.1395	1.2142	1.4468	1.2952	1.5849	1.6353	2.1214	1.6562	2.1440
[4]	1.6656	2.1061	1.5932	2.0518	1.2246	1.4592	1.3051	1.5965	1.5883	2.0467	1.5690	2.0143
[5]	2.1526	2.8298	2.1217	2.8244	1.4289	1.7026	1.5625	1.9313	2.1204	2.8675	2.0944	2.7809
[6]	2.0219	2.8139	1.9415	2.5615	1.3863	1.6519	1.5039	1.8530	1.9146	2.5256	1.9121	2.5117
[7]	1.8253	2.3360	1.7234	2.2392	1.2655	1.5079	1.3562	1.6627	1.7120	2.2377	1.7132	2.2203
[8]	1.7290	2.2818	1.6054	2.0685	1.2395	1.4770	1.3220	1.6177	1.6076	2.0631	1.5954	2.0493
[9]	1.6816	2.1968	1.6431	2.1214	1.2488	1.4880	1.3343	1.6339	1.6136	2.0929	1.6358	2.1071
[10]	1.8064	2.2591	1.6754	2.1675	1.2566	1.4973	1.3445	1.6474	1.6653	2.1710	1.6636	2.1482
[11]	1.7245	2.2904	1.6782	2.1775	1.2430	1.4812	1.3285	1.6270	1.6886	2.2053	1.6426	2.1253
[12]	1.6759	2.1434	1.6449	2.1252	1.2481	1.4872	1.3333	1.6326	1.6374	2.1200	1.6348	2.1033
[13]	1.8299	2.4993	1.7724	2.3044	1.3030	1.5526	1.4010	1.7199	1.7660	2.2984	1.7672	2.2909
[14]	1.6007	2.0857	1.6130	2.0789	1.2401	1.4776	1.3232	1.6194	1.5938	2.0671	1.5858	2.0396
[15]	1.7540	2.2729	1.6768	2.1690	1.2731	1.5170	1.3625	1.6695	1.6698	2.1834	1.6496	2.1262
[16]	1.7848	2.3377	1.7207	2.2350	1.2532	1.4933	1.3429	1.6464	1.6982	2.2097	1.7251	2.2365
[17]	1.7424	2.1501	1.6597	2.1438	1.2722	1.5159	1.3607	1.6669	1.6277	2.1042	1.6401	2.1126

Table 3, continued.

	C_1		C_2		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[18]	1.7336	2.1373	1.6528	2.1345	1.2618	1.5035	1.3490	1.6523	1.6297	2.1138	1.6378	2.1099
[19]	1.0242	1.2681	1.0099	1.2531	0.8758	1.0436	0.9047	1.0926	1.0153	1.2497	1.0011	2.2410
[20]	1.0137	1.2284	0.9834	1.2145	0.8717	1.0387	0.8998	1.0864	0.9957	1.2302	0.9712	1.1978
[21]	1.8246	2.3465	1.8066	2.3495	1.3169	1.5692	1.4194	1.7442	1.8067	2.3370	1.8018	2.3372
[22]	1.6455	2.0421	1.6180	2.0857	1.2377	1.4748	1.3211	1.6170	1.6001	2.0816	1.5875	2.0391
[23]	1.0462	1.2845	1.0328	1.2825	0.8884	1.0586	0.9189	1.1104	1.0393	1.2960	1.0311	1.2787
[24]	1.0103	1.2819	0.9854	1.2171	0.8753	1.0430	0.9036	1.0911	0.9800	1.2079	0.9812	1.2099
[25]	0.7842	0.9543	0.7775	0.9509	0.7016	0.8360	0.7165	0.8613	0.7766	0.9502	0.7754	0.9475
[26]	0.7846	0.9490	0.7714	0.9407	0.7079	0.8435	0.7229	0.8691	0.7677	0.9354	0.7671	0.9342
[27]	0.6895	0.8386	0.6832	0.8310	0.6328	0.7540	0.6436	0.7723	0.6820	0.8351	0.6820	0.8275
[28]	0.6546	0.8045	0.6550	0.7944	0.6162	0.7343	0.6261	0.7510	0.6526	0.7914	0.6561	0.7941
[29]	0.6009	0.7334	0.5902	0.7144	0.5567	0.6634	0.5640	0.6758	0.5945	0.7184	0.5879	0.7109
[30]	0.5796	0.7047	0.5780	0.6982	0.5513	0.6569	0.5583	0.6688	0.5752	0.6973	0.5761	0.6951
[31]	0.7042	0.8616	0.7249	0.8823	0.6713	0.7999	0.6842	0.8217	0.7312	0.8881	0.7259	0.8817
[32]	0.6482	0.7763	0.6563	0.7960	0.6176	0.7359	0.6275	0.7526	0.6639	0.8022	0.6546	0.7929
[33]	0.4067	0.4736	0.4067	0.4884	0.3951	0.4708	0.3977	0.4752	0.4043	0.4892	0.4047	0.4859
[34]	0.3985	0.4815	0.3992	0.4789	0.3897	0.4644	0.3922	0.4686	0.4014	0.4818	0.3968	0.4754

Table 4. Coverage to Length Ratio (CLR) of confidence intervals

	C_1		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[1]	0.3873	0.3026	0.5908	0.5180	0.5838	0.4988	0.4471	0.3582	0.4401	0.3497
[2]	0.4119	0.3240	0.6046	0.5280	0.5892	0.5044	0.4705	0.3752	0.4667	0.3737
[3]	0.4925	0.3854	0.6860	0.6002	0.6770	0.5840	0.5484	0.4460	0.5449	0.4433
[4]	0.4977	0.4125	0.6845	0.6009	0.6754	0.5816	0.5675	0.4655	0.5759	0.4713
[5]	0.3811	0.3007	0.5711	0.4992	0.5623	0.4766	0.4234	0.3305	0.4305	0.3423
[6]	0.4130	0.3074	0.6033	0.5270	0.5871	0.5022	0.4689	0.3773	0.4680	0.3770

Table 4, continued.

	C1		C3		C4		C5		C6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[7]	0.4489	0.3674	0.6541	0.5730	0.6442	0.5551	0.5256	0.4255	0.5287	0.4300
[8]	0.4835	0.3807	0.6791	0.5943	0.6682	0.5739	0.5602	0.4618	0.5640	0.4627
[9]	0.4977	0.3953	0.6725	0.5863	0.6591	0.5655	0.5609	0.4552	0.5493	0.4508
[10]	0.4625	0.3832	0.6656	0.5788	0.6530	0.5649	0.5372	0.4356	0.5397	0.4425
[11]	0.4791	0.3753	0.6684	0.5863	0.6641	0.5700	0.5343	0.4305	0.5510	0.4478
[12]	0.4985	0.4036	0.6737	0.5916	0.6672	0.5718	0.5505	0.4488	0.5465	0.4510
[13]	0.4546	0.3451	0.6460	0.5636	0.6311	0.5435	0.5097	0.4135	0.5073	0.4156
[14]	0.5294	0.4222	0.6833	0.5979	0.6700	0.5769	0.5622	0.4583	0.5659	0.4639
[15]	0.4771	0.3845	0.6589	0.5746	0.6423	0.5541	0.5374	0.4364	0.5461	0.4460
[16]	0.4655	0.3693	0.6633	0.5802	0.6565	0.5624	0.5328	0.4314	0.5188	0.4246
[17]	0.4839	0.4057	0.6675	0.5817	0.6519	0.5577	0.5532	0.4517	0.5487	0.4480
[18]	0.4798	0.4066	0.6649	0.5824	0.6514	0.5604	0.5488	0.4489	0.5494	0.4503
[19]	0.8389	0.7061	1.0037	0.8741	0.9840	0.8618	0.8825	0.7610	0.8941	0.7621
[20]	0.8563	0.7382	0.9987	0.8759	0.9851	0.8614	0.9041	0.7745	0.9224	0.7863
[21]	0.4492	0.3641	0.6287	0.5484	0.6174	0.5323	0.4976	0.4063	0.4982	0.4061
[22]	0.5088	0.4270	0.6787	0.5907	0.6634	0.5754	0.5607	0.4584	0.5650	0.4661
[23]	0.8258	0.7063	0.9721	0.8496	0.9640	0.8433	0.8637	0.7323	0.8709	0.7393
[24]	0.8685	0.7121	1.0031	0.8756	0.9920	0.8646	0.9085	0.7836	0.9183	0.7862
[25]	1.1112	0.9599	1.2500	1.0955	1.2488	1.0951	1.1493	0.9943	1.1614	0.9990
[26]	1.1244	0.9705	1.2499	1.0957	1.2444	1.0935	1.1733	1.0148	1.1651	1.0135
[27]	1.2827	1.1026	1.3891	1.2218	1.3850	1.2156	1.3199	1.1378	1.3153	1.1447
[28]	1.3523	1.1562	1.4411	1.2654	1.4298	1.2610	1.3920	1.2045	1.3639	1.1959
[29]	1.4818	1.2648	1.6077	1.4109	1.5996	1.4049	1.5180	1.3177	1.5220	1.3368
[30]	1.5349	1.3265	1.6176	1.4270	1.6088	1.4181	1.5722	1.3668	1.5584	1.3635
[31]	1.2667	1.0822	1.3192	1.1561	1.3099	1.1513	1.2319	1.0727	1.2354	1.0759
[32]	1.3675	1.1988	1.4372	1.2687	1.4298	1.2594	1.3651	1.1885	1.3684	1.1954
[33]	2.1957	1.9793	2.2622	1.9983	2.2625	1.9895	2.2158	1.9311	2.2351	1.9614
[34]	2.2394	1.9556	2.3120	2.0353	2.3014	2.0320	2.2291	1.9611	2.2857	2.0076

We denote by C_1 the CI proposed by Balakrishnan and Asgharzadeh (2005), by C_2 the CI proposed Wang (2009), by C_3 the CI based on the MLE obtained by the EM algorithm, by C_4 the CI based on the log-transformed MLE, by C_5 the CI based on pivotal quantity, and by C_6 the GCI. Coverage probabilities of the CIs for various censoring schemes are displayed in Table 2. Coverage probabilities of C_1 are also displayed in the same table. Coverage probabilities for C_2 are not provided by Wang (2009). Lengths of CIs for the various censoring schemes are given in Table 3. For comparison, lengths of C_1 and C_2 are given in the same table.

For effective comparison of CIs, we compute coverage to length ratio (CLR). CLR for C_1, C_3, C_4, C_5, and C_6 are given in Table 4. It is clear that the CIs having a higher value of CLR are preferred.

Conclusion

Coverage probabilities of C_3, C_4, C_5, and C_6 are better than coverage probabilities of C_1. Comparing coverage probabilities of all four CIs, C_5 and C_6 show the best performance. For small and large sample sizes (n) and the smallest effective sample size (m), C_5 and C_6 show good coverage probability. For large sample sizes, C_3, C_4, C_5, and C_6 show good performance. As n and m increase, coverage probability of C_3 and C_4 increases rapidly as compared to C_5 and C_6. C_6 has higher coverage probability for conventional censoring schemes than progressive censoring schemes, but C_3 and C_4 show higher coverage probability for progressive censoring schemes than conventional censoring schemes.

C_3 has smaller length than the lengths of C_1 and C_2. The MLE by the EM algorithm provides the shortest length CI among all five CIs. For large sample sizes, the length of C_6 approaches the length of C_3. Lengths of all CIs decrease as n and m increase. Lengths of CIs based on progressive censoring schemes are smaller than lengths of CIs based on conventional censoring schemes. There is a minor difference among lengths of C_3, C_4, C_5, and C_6 for large sample sizes. According to the CLR, C_3 is the best among the four CIs for small sample sizes. C_4, C_5, and C_6 also show higher CLR than the CLR of C_1. CLRs of CIs based on progressive censoring schemes are better than CLRs of CIs based on conventional censoring.

Acknowledgements

The first author wishes to thank the University Grants Commission, New Delhi, India for providing fellowship under the Faculty Improvement Programme to carry out this research.

References

Balakrishnan, N. (2007). Progressive censoring methodology: An appraisal (with discussion). *Test, 16*(2), 211-296.doi: 10.1007/s11749-007-0061-y

Balakrishnan, N., & Aggarwala, R. (2000). *Progressive censoring: Theory, methods, and applications*. Boston, MA: Birkhauser.doi: 10.1007/978-1-4612-1334-5

Balakrishnan, N., & Asgharzadeh, A. (2005). Inference for the scaled half-logistic distribution based on progressively Type-II censored samples. *Communications in Statistics – Theory and Methods, 34*(1), 73-87.doi: 10.1081/sta-200045814

Balakrishnan, N., & Chan, P.S. (1992). Estimation for the scaled half logistic distribution under Type-II censoring.*Computational Staistics& Data Analysis, 13*(2), 123-141.doi: 10.1016/0167-9473(92)90001-v

Balakrishnan, N., Kannan, N., Lin, C.T., & Ng, H. K. T. (2003). Point and interval estimation for Gaussian distribution, based on progressively Type-II censored samples.*IEEE Transactions on Reliability, 52*(1), 90-95. doi: 10.1109/tr.2002.805786

Balakrishnan, N., Kannan, N., Lin, C.T., & Wu, S. J. S. (2004). Inference for the extreme value distribution under progressive Type-II Censoring.*Journal of Statistical Computationand Simulation, 74*(1), 25-45.doi: 10.1080/0094965031000105881

Balakrishnan, N., &Puthenpura, S. (1986). Best linear unbiased estimation of location and scale parameters of the half logistic distribution.*Journal of Statistical Computation and Simulation, 25*(3-4), 193-204.doi: 10.1080/00949658608810932

Balakrishnan, N., & Sandhu, R. A. (1995). A simple simulation algorithm for generating progressive Type-II censored samples.*The American Statistician, 49*(2), 229-230.doi: 10.2307/2684646

Balakrishnan, N., & Wong, K. H. T. (1991). Approximate MLEs for the location and scaled parameters of the half logistic distribution with Type-II right censoring.*IEEE Transactions on Reliability, 40*(2), 140-145. doi: 10.1109/24.87114

Cohen, A. C. (1963). Progressively censored samples in life testing.*Technometrics, 5*(3), 327-329.doi: 10.2307/1266337

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm.*Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38. Available from: http://www.jstor.org/stable/2984875

Ghitany, M. E., Alqallaf, F., & Balakrishnan, N. (2014). On the likelihood estimation of the parameters of Gompertz distribution based on complete and progressively Type-II censored samples. *Journal of Statistical Computation and Simulation, 84*(8), 1803-1812.doi: 10.1080/00949655.2013.766738

Gulati, S., & Mi, J. (2006). Testing for scale families using total variation distance.*Journal of Statistical Computation and Simulation, 76*(9), 773-792.doi: 10.1080/10629360500282080

Jang, D. H., Park, J., & Kim, C. (2011). Estimation of the scale parameter of the half-logistic distribution with multiply type II censored sample.*Journal of the Korean Statistical Society, 40*(3), 291-301.doi: 10.1016/j.jkss.2010.12.001

Kim, C., & Han, K. (2010). Estimation of the scale parameter of the half-logistic distribution under progressively typeII censored sample.*Statistical Papers, 51*(2), 375-387.doi: 10.1007/s00362-009-0197-9

Lawless, J. F. (1982).*Statistical models and methods for lifetime data*.New York, NY: John Wiley and Sons.

Lin, C.-T., & Balakrishnan, N. (2011). Asymptotic properties of maximum likelihood estimators based on progressively Type-II censoring.*Metrika, 74*(3), 349-360.doi: 10.1007/s00184-010-0306-8

Louis, T. A. (1982). Finding the observed information matrix using the EM algorithm.*Journal of the Royal Statistical Society. Series B (Methodological), 44*(2), 226-233. Available from: http://www.jstor.org/stable/2345828

Mann, N. R. (1969). Exact three-order-statistic confidence bounds on reliable life for a Weibull model with progressive censoring. Journal of the American Statistical Association, 64(325), 306-315.doi: 10.2307/2283740

Mann, N. R. (1971). Best linear invariant estimation for Weibull parameters under progressive censoring.*Technometrics, 13*(3), 521-533.doi: 10.2307/1267165

McLachlan, G. J., & Krishnan, T. (1997).*The EM algorithm and extensions*.New York, NY: John Wiley and Sons.

Meeker, W. Q., & Escobar, L. A. (1998).*Statistical methods for reliability data*.New York, NY: John Wiley and Sons.

Ng, H. K. T. (2005). Parameter estimation for a modified Weibull distribution for progressively Type-II censored samples.*IEEE Transactions on Reliability, 54*(3), 374-380.doi: 10.1109/tr.2005.853036

Ng, H. K. T., Kundu, D., & Balakrishnan, N. (2006). Point and interval estimation for the two-parameter Birnbaum-Saunders distribution based on progressively Type-II censored data.*Computational Statistics & Data Analysis, 50*(11), 3222-3242.doi: 10.1016/j.csda.2005.06.002

Potdar, K. G., & Shirke, D. T. (2013). Reliability estimation for the distribution of a *k*-unit parallel system with Rayleigh distribution as the component life distribution.*International Journal of Engineering Research and Technology, 2*(8), 2362-2371.

Potdar, K. G., & Shirke, D. T. (2014). Inference for the scale parameter of lifetime distribution of *k*-unit parallel system based on progressively censored data. *Journal of Statistical Computation and Simulation, 84*(1), 171-185.doi: 10.1080/00949655.2012.700314

Rastogi, M. K., & Tripathi, Y. M. (2014). Parameter and reliability estimation for an exponentiated half-logistic distribution under progressive type II censoring.*Journal of Statistical Computation and Simulation, 84*(8), 1711-1727. doi: 10.1080/00949655.2012.762366

Sultan, K. S., Alsadat, N. H., & Kundu, D. (2014). Bayesian and maximum likelihood estimations of the inverse Weibull parameters under progressive type-II censoring. Journal of Statistical Computation and Simulation, 84(10), 2248-2265.doi: 10.1080/00949655.2013.788652

Wang, B. (2009). Interval estimation for the scaled half logistic distribution under progressive Type-II censoring.*Communicationsin Statistics – Theory and Methods, 38*(3), 364-371.doi: 10.1080/03610920802213681

Weerahandi, S. (1993). Generalized confidence intervals. Journal of the American Statistical Association, 88(423), 899-905.doi: 10.2307/2290779

Appendix A. Illustrative Examples

Numeric Example

Balakrishnan and Asgharzadeh (2005) gave simulated sample of size $n = 50$ from the half-logistic distribution with scale parameter $\lambda = 25$. This complete sample is

1.7110, 2.0024, 2.3963, 3.9034, 4.6412, 6.4002, 6.7956, 8.5646, 8.6428, 8.8354, 9.3518, 9.7358, 10.5080, 10.5095, 11.8015, 12.8005, 16.3451, 16.9938, 17.2101, 18.5384, 20.3508, 21.1838, 22.1529, 22.4062, 22.4381, 23.0369, 25.8435, 27.0574, 27.1237, 29.0360, 30.6449, 32.5713, 33.6688, 40.3890, 45.4092, 46.4756, 49.8833, 51.1798, 53.0397, 53.8135, 64.9315, 66.1807, 69.9004, 75.2674, 75.4427, 75.7291, 76.1571, 89.5827, 99.8525, 134.6488.

Balakrishnan and Asgharzadeh (2005) and Wang (2009) derived CIs for this complete sample and the censored sample. We also derive CIs by using the MLE obtained by the EM algorithm, and the CIs based on pivot and generalized pivot. In Table 5, we consider two cases suggested by Wang (2009). Also we use the censoring schemes and samples given by Wang (2009) and derive 90% and 95% CIs and their lengths. For comparison, we display CIs and their lengths as stated by Wang (2009).

Table 5. Confidence interval and its length for illustrative example: $n = 50$, $\lambda = 25$

	C_2		C_3	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(24.49, 42.97)	(23.37, 45.72)	(22.76, 40.26)	(21.08, 41.94)
(25*1)	18.48	22.35	17.50	20.86
Case 2	(20.93, 34.82)	(20.05, 36.81)	(19.95, 33.28)	(18.67, 34.56)
(28*0, 10,10)	13.89	16.76	13.33	15.89

	C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(24.52, 42.94)	(23.38, 45.67)	(24.05, 42.82)	(23.18, 45.66)
(25*1)	18.42	22.29	18.77	22.48
Case 2	(21.21, 35.21)	(20.31, 37.23)	(21.42, 34.93)	(20.31, 37.24)
(28*0, 10,10)	14.00	16.92	13.51	16.93

Note: For Case 1, Sr. No. is 1 and $m = 25$. For Case 2, Sr. No. is 2 and $m = 30$.

Table 6. Confidence interval and its length for illustrative example: $n = 50$, $\lambda = 25$

	C_1		C_3	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(19.81, 29.53)	(18.90, 30.45)	(19.88, 29.48)	(18.96, 30.40)
(50*0)	9.72	11.55	9.6	11.44
Case 2	(20.78, 32.12)	(19.72, 33.18)	(18.88, 29.21)	(17.89, 30.20)
(39*0, 10)	11.34	13.46	10.33	12.31
Case 3	(18.66, 31.16)	(17.48, 32.34)	(15.92, 26.62)	(14.89, 27.65)
(29*0, 20)	12.5	14.86	10.7	12.76

	C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(20.59, 30.37)	(19.85, 31.60)	(20.55, 30.26)	(19.92, 31.28)
(50*0)	9.78	11.75	9.71	11.36
Case 2	(19.68, 30.38)	(18.94, 31.81)	(19.53, 30.07)	(18.95, 31.47)
(39*0, 10)	10.7	12.87	10.54	12.52
Case 3	(16.95, 28.23)	(16.23, 29.80)	(16.90, 28.20)	(16.06, 29.92)
(29*0, 20)	11.28	13.57	11.3	13.86

Note: For Case 1, Sr. No. is 1 and $m = 50$. For Case 2, Sr. No. is 2 and $m = 40$. For Case 3, Sr. No. is 3 and $m = 30$.

Balakrishnan and Asgharzadeh (2005) considered three cases, ($n = 50$, $m = 50$), ($n = 50$, $m = 40$), and ($n = 50$, $m = 30$). They used progressive and conventional Type-II censored samples but have not provided samples. To compare the proposed CIs with the CI proposed by Balakrishnan and Asgharzadeh (2005), we considered conventional censored and complete samples considered by Balakrishnan and Asgharzadeh (2005). We obtained 90% and 95% CIs for these schemes. In Table 6, 90% and 95% CIs and their lengths are displayed. Also, the CIs and their length proposed by Balakrishnan and Asgharzadeh (2005) are displayed.

Observe that in the illustrated example, C_3 has shorter length than the lengths of C_1, C_2 and C_5. C_6 has shorter length than that of C_1.

Real Data Example

Lawless (1982) presented real data which represented failure times for a specific type of electrical insulation that was subjected to a continuously increasing voltage stress.

12.3, 21.8, 24.4, 28.6, 43.2, 46.9, 70.7, 75.3, 95.5, 98.1, 138.6, 151.9.

Table 7. Confidence interval and its length for real data: $n = 12$, $\lambda = 50.50$ (BLUE)

	C_3		C_4	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(28.59, 66.24)	(24.98, 69.85)	(31.88, 70.53)	(29.54, 76.10)
(12*0)	37.65	44.87	38.65	46.56
Case 2	(25.55, 73.70)	(20.94, 78.31)	(30.55, 80.61)	(27.84, 88.46)
(7*0, 4)	48.15	57.37	50.06	60.62
Case 3	(23.35, 68.29)	(19.05, 72.59)	(28.06, 74.82)	(25.54, 82.19)
(4, 7*0)	44.94	53.54	46.74	56.65

	C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(33.37, 75.18)	(31.19, 82.30)	(33.65, 73.96)	(31.88, 83.36)
(12*0)	41.81	51.11	40.31	51.48
Case 2	(33.13, 90.13)	(30.73, 101.89)	(32.60, 86.50)	(30.13, 94.26)
(7*0, 4)	57	71.16	53.9	64.13
Case 3	(30.14, 82.01)	(27.78, 92.25)	(30.55, 83.15)	(27.58, 92.42)
(4, 7*0)	51.87	64.47	52.6	64.84

Note: For Case 1, Sr. No. is 1 and $m = 12$. For Case 2, Sr. No. is 2 and $m = 8$. For Case 3, Sr. No. is 3 and $m = 8$.

The half-logistic distribution fits the data extremely well (Balakrishnan & Chan, 1992). This dataset was used with two censoring schemes, (7*0, 4) and (4, 7*0), and complete data, and the CI is constructed based on the MLE, log-MLE, pivot, and generalized pivot. These 90% and 95% CIs and their lengths are presented in Table 7. Observe that, for real data, C_3 has shorter length than C_4, C_5 and C_6.

The EM algorithm approach works well for small sample size n and the smallest effective sample size m. Overall, the proposed CIs perform better than the CIs proposed by Balakrishnan and Asgharzadeh (2005) and Wang (2009). The proposed CIs are superior to the other two CIs with regard to the length and the coverage probability.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 350-363.
doi: 10.22237/jmasm/1493597940

ISSN 1538 − 9472

A New Estimator for the Pickands Dependence Function

Marta Ferreira
Universidade do Minho
Braga, Portugal

The Pickands dependence function characterizes an extreme value copula, a useful tool in the modeling of multivariate extremes. A new estimator is presented along with its convergence properties and performance through simulation.

Keywords: Extreme value copula; tail dependence; nonparametric estimation

Introduction

Tail dependence is an important issue in several areas like finance, environment, engineering, among others, given the concern on the impact of the occurrence of joint extreme events. The copula concept provides a margin-free tool to describe the dependence structure of a random vector. Focusing on the bivariate case from now on, given a random pair (X, Y) with joint distribution function (df) H, then it may be represented as

$$\mathrm{H}(x,y)=\mathrm{C}\big(\mathrm{F}(x),\mathrm{G}(y)\big)$$

for all x, y $\in \mathbb{R}$, where F and G are the marginal df's of X and Y, respectively. We always assume that F and G are continuous and thus copula C is unique (Sklar, 1959). Considering U = F(X) and V = G(Y), we may also write

$$\mathrm{C}(u,v)=\mathrm{P}\big(U\le u,V\le v\big)$$

for all $u, v \in [0, 1]$. Extreme-value copulas arise in the limit of an increasing sample length of copulas of componentwise maxima of independent or strongly

Marta Ferreira is a researcher in the Centro de Matemática. Email them at: msferreira@math.uminho.pt.

mixing stationary sequences (Deheuvels, 1984; Hsing, 1989). Extreme-value copulas are completely determined by the Pickands dependence function, A: [0, 1] → [1/2, 1], which is convex and satisfies $t \vee (1-t) \leq \mathrm{A}(t) \leq 1$, $\forall t \in [0, 1]$, where $x \vee y = \max(x, y)$. More precisely, for all $0 \leq u, v \leq 1$,

$$\mathrm{C}(u,v) = \exp\left(\log(uv)\, \mathrm{A}\left(\frac{\log(v)}{\log(uv)} \right) \right) \tag{1}$$

Modeling applications of extreme-value copulas can be seen in Tawn (1988), Ghoudi, Khoudraji, and Rivest (1998), Frees and Valdez (1998), Coles, Heffernan, and Tawn (1999), Cebrian, Denuit, and Lambert (2003), McNeil, Frey, and Embrechts (2005), Salvadori, De Michele, Kottegoda, and Rosso (2007), amongst others. For instance, in volatile and bear markets, a dependence measure often used in lieu of Pearson's correlation to account for extreme events dependence is the so-called tail dependence coefficient (TDC) introduced in Sibuya (1960), usually denoted λ, which corresponds to $2(1 - \mathrm{A}(0.5))$. The TDC ranges in [0, 1]. The null boundary case corresponds to asymptotic tail independence, a very important topic in the statistics of extremes. Indeed, this case may not correspond to perfect independence but to a “residual" one that must be taken into account in order to avoid misleading risk estimates. See, e.g., Beirlant, Goegebeur, Segers, and Teugels (2004) and references therein.

Other representations than (1) may be considered, e.g., based on the stable tail dependence function, l: $[0, \infty)^2 \to [0,\infty)$, which is convex, homogeneous of order one (i.e., $l(\alpha x, \alpha y) = \alpha l(x, y)$ for $\alpha > 0$), satisfies $x \vee y \leq l(x, y) \leq x + y$, $\forall x, y \geq 0$, and $l(x, y) = (x + y)\mathrm{A}(y/(x + y))$, thus leading to

$$\mathrm{C}(u,v) = \exp\left(-l\left(-\log(u), -\log(v) \right) \right)$$

Representation (1) can also be formulated as

$$\mathrm{C}\left(w^{1-t}, w^{t} \right) = w^{\mathrm{A}(t)}$$

and thus, as well,

$$\mathrm{C}\left(w^{1-t}, w^{t} \right) = w^{l(1-t,t)}$$

Therefore, statistical inference on a bivariate extreme-value copula can be reduced to the estimation of a univariate Pickands dependence function (or a bivariate stable tail dependence function, although they are related).

Several parametric and non-parametric estimators of the Pickands dependence function are found in the literature. A wide survey on this topic is presented in Beirlant et al. (2004). Nonparametric estimation has been essentially based on the Pickands estimator (Pickands, 1981) and on the Capéraà-Fougères-Genest (CFG) estimator (Capéraà, Fougères, & Genest, 1997). Further modifications of the former can be seen in Deheuvels (1991) and Hall and Tajvidi (2000), while the latter can be found in Jiménez, Villa-Diharce, and Flores (2001), Zhang, Wells, and Peng (2008), and Gudendorf and Segers (2011); for both, see Segers (2007). All these approaches assume known margins, which is rather unrealistic in practice. Nonparametric versions of the Pickands and CFG estimators based on unknown margins are addressed in Abdous and Ghoudi (2005), Genest and Segers (2009), and Gudendorf and Segers (2012).

Pickands Dependence Function: Estimators and Properties

Let (X, Y) be a random pair with joint df H and continuous marginal df's F and G, respectively, such that, $U = \mathrm{F}(X)$ and $V = \mathrm{G}(Y)$. Let C be a bivariate extreme-value copula, i.e. of the form (1), characterizing the dependence between X and Y. Thus C is the df of the random pair (U, V).

Consider $S = -\log(U)$, $T = -\log(V)$ and

$$\xi(0) = \frac{S}{1-t} \wedge \frac{T}{t}, 0 < t < 1$$

with $\xi(0) = S$ and $\xi(1) = T$. The random variables (rv's) S and T are Exponential with unit mean value and $\xi(t)$ is also exponentially distributed with mean values

$$\mathrm{E}\left(\xi(t)\right) = \frac{1}{\mathrm{A}(t)} \quad \text{and} \quad \mathrm{E}\left(\log\left(\xi(t)\right)\right) = -\log\left(\mathrm{A}(t)\right) - \gamma$$

where γ denotes the Euler's constant $\int_0^\infty \log(x)\mathrm{e}^{-x}dx \approx 0.577$. These relations are the bases of, respectively, the Pickands and the CFG estimators by considering the empirical counterparts. More precisely, for a random sample $(X_1, Y_1),\ldots, (X_n, Y_n)$

distributed as (X, Y) such that $U_i = \mathrm{F}(X_i)$ and $V_i = \mathrm{G}(Y_i)$, $S_i = -\log(U_i) = \xi_i(0)$, $T_i = -\log(V_i) = \xi_i(1)$ for all $i = 1,\ldots, n$, with

$$\xi_i(t) = \frac{S_i}{1-t} \wedge \frac{T_i}{t}, 0 < t < 1$$

we have

$$\frac{1}{\mathrm{A}_n^{\mathrm{P}}(t)} = \frac{1}{n}\sum_{i=1}^{n} \xi_i(t)$$

and

$$\log\left(\mathrm{A}_n^{\mathrm{CFG}}(t)\right) = -\gamma - \frac{1}{n}\sum_{i=1}^{n} \log\left(\xi_i(t)\right)$$

Whenever the margins F and G are unknown, the natural approach is to consider the respective marginal empirical df's F_n and G_n and take

$$\hat{U}_i = \frac{n\mathrm{F}_n(X_i)}{n+1} = \frac{1}{n+1}\sum_{j=1}^{n} \mathbb{I}_{\{X_j \le X_i\}} \quad \text{and} \quad \hat{V}_i = \frac{n\mathrm{G}_n(Y_i)}{n+1} = \frac{1}{n+1}\sum_{j=1}^{n} \mathbb{I}_{\{Y_j \le Y_i\}} \qquad (2)$$

where $\mathbb{I}$ is the indicator function. The replacement of U_i and V_i everywhere in the expressions above by, respectively, $\hat{U}_i$ and $\hat{V}_i$, leads now to

$$\frac{1}{\hat{\mathrm{A}}_n^{\mathrm{P}}(t)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_i(t)$$

and

$$\log\left(\hat{\mathrm{A}}_n^{\mathrm{CFG}}(t)\right) = -\gamma - \frac{1}{n}\sum_{i=1}^{n} \log\left(\hat{\xi}_i(t)\right)$$

In order to satisfy the endpoint constraints $\mathrm{A}(0) = \mathrm{A}(1) = 1$, endpoint corrected versions were considered, namely,

$$\frac{1}{\mathrm{A}^{\mathrm{P}}_{n,c}(t)} = \frac{1}{\mathrm{A}^{\mathrm{P}}_{n}(t)} - (1-t)\left(\frac{1}{\mathrm{A}^{\mathrm{P}}_{n}(0)} - 1\right) - t\left(\frac{1}{\mathrm{A}^{\mathrm{P}}_{n}(1)} - 1\right)$$

and

$$\log\left(\mathrm{A}^{\mathrm{CFG}}_{n,c}(t)\right) = \log\left(\mathrm{A}^{\mathrm{CFG}}_{n}(t)\right) - (1-t)\log\left(\mathrm{A}^{\mathrm{CFG}}_{n}(0)\right) - t\log\left(\mathrm{A}^{\mathrm{CFG}}_{n}(1)\right)$$

Further developments on this topic can be found in Segers (2007). Similar procedures can be applied to the case of unknown marginal estimators and thus derive $\hat{\mathrm{A}}^{\mathrm{P}}_{n,c}(t)$ and $\hat{\mathrm{A}}^{\mathrm{CFG}}_{n,c}(t)$, although they are asymptotically equivalent to the respective uncorrected $\hat{\mathrm{A}}^{\mathrm{P}}_{n}(t)$ and $\hat{\mathrm{A}}^{\mathrm{CFG}}_{n}(t)$, as shown in Genest and Segers (2009). Another correction of the Pickands estimator based on Hall and Tajvidi (2000) is to consider

$$\frac{1}{\hat{\mathrm{A}}^{\mathrm{HT}}_{n}(t)} = \frac{1}{n}\sum_{i=1}^{n} \overline{\xi}_i(t)$$

with

$$\overline{\xi}_i(t) = \frac{\overline{S}_i}{1-t} \wedge \frac{\overline{T}_i}{t}$$

where $\overline{S}_i = n\hat{S}_i \big/ \left(\hat{S}_1 + \ldots + \hat{S}_n\right)$ and $\overline{T}_i = n\hat{T}_i \big/ \left(\hat{T}_1 + \ldots + \hat{T}_n\right)$, $\hat{S}_i = -\log\left(\hat{U}_i\right) = \hat{\xi}_i(0)$, $\hat{T}_i = -\log\left(\hat{V}_i\right) = \hat{\xi}_i(1)$, $i = 1,\ldots, n$. We have $\hat{\mathrm{A}}^{\mathrm{HT}}_{n}(0) = \hat{\mathrm{A}}^{\mathrm{HT}}_{n}(1) = 1$ and also $\hat{\mathrm{A}}^{\mathrm{HT}}_{n}(t) \geq t \vee (1-t)$ for all $0 \leq t \leq 1$. Relation $\hat{\mathrm{A}}^{\mathrm{HT}}_{n}(t) = \hat{\mathrm{A}}^{\mathrm{P}}_{n}(t) \big/ \hat{\mathrm{A}}^{\mathrm{P}}_{n}(0)$ means that $\hat{\mathrm{A}}^{\mathrm{HT}}_{n}(t)$ and $\hat{\mathrm{A}}^{\mathrm{P}}_{n}(t)$ are asymptotically equivalent, too.

The asymptotic properties of estimators $\hat{\mathrm{A}}^{\mathrm{P}}_{n}(t)$ and $\hat{\mathrm{A}}^{\mathrm{CFG}}_{n}(t)$, derived in Genest and Segers (2009), are based on the empirical copula

$$\hat{\mathrm{C}}_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\left\{\hat{U}_i \leq u, \hat{V}_i \leq v\right\}}, \ \forall u,v \in [0,1]$$

More precisely, Genest and Segers' Lemma 3.1 states that, for all $t \in [0, 1]$,

$$\sqrt{n}\left(\frac{1}{\hat{\mathrm{A}}_n^{\mathrm{P}}(t)}-\frac{1}{\mathrm{A}(t)}\right)=\frac{\int_0^1 \mathbb{C}_n\left(u^{1-t},u^t\right)du}{u} \tag{3}$$

$$\sqrt{n}\left(\log\hat{\mathrm{A}}_n^{\mathrm{CFG}}(t)-\log\mathrm{A}(t)\right)=\frac{\int_0^1 \mathbb{C}_n\left(u^{1-t},u^t\right)du}{u\log u} \tag{4}$$

where $\mathbb{C}_n$ is the empirical copula process $\sqrt{n}\left(\hat{\mathrm{C}}_n-\mathrm{C}\right)$. Now consider $\alpha_n=\sqrt{n}\left(\mathrm{C}_n-\mathrm{C}\right)$, with

$$\mathrm{C}_n(u,v)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{U_i\le u,V_i\le v\}}\ ,\ \forall u,v\in[0,1]$$

The classical theory of empirical processes states that the weak limit α of the process $\alpha_n=\sqrt{n}\left(\mathrm{C}_n-\mathrm{C}\right)$ is a centered Gaussian process with covariance

$$\mathrm{cov}\left(\alpha(u,v),\alpha(u',v')\right)=\mathrm{C}(u\wedge v,u'\wedge v')-\mathrm{C}(u,v)\mathrm{C}(u',v'),\ \forall u,v,u',v'\in[0,1]$$

The weak limit $\mathbb{C}$ of the process $\mathbb{C}_n=\sqrt{n}\left(\hat{\mathrm{C}}_n-\mathrm{C}\right)$ is closely related to α, namely,

$$\mathbb{C}(u,v)=\alpha(u,v)-\frac{\partial\mathrm{C}(u,v)}{\partial u}\alpha(u,1)-\frac{\partial\mathrm{C}(u,v)}{\partial v}\alpha(1,v),\ \forall(u,v)\in[0,1]^2$$

If A is twice continuously differentiable on (0, 1) and $\sup_{\{0<t<1\}}t(1-t)\mathrm{A}''(t)<\infty$, then the following weak convergence results hold, as $n\to\infty$, in the space $\mathcal{C}([0,1])$ of continuous and real-valued functions on [0, 1] equipped with the topology of uniform convergence:

$$\mathbb{A}_n^{\mathrm{P}}=\sqrt{n}\left(\hat{\mathrm{A}}_n^{\mathrm{P}}(t)-\mathrm{A}(t)\right)\xrightarrow{w}\mathbb{A}^{\mathrm{P}}(t)=-\mathrm{A}^2(t)\int_0^1\frac{\mathbb{C}\left(u^{1-t},u^t\right)du}{u} \tag{5}$$

and

$$\mathbb{A}_n^{\mathrm{CFG}} = \sqrt{n}\left(\hat{\mathrm{A}}_n^{\mathrm{CFG}}(t) - \mathrm{A}(t)\right) \xrightarrow{w} \mathbb{A}^{\mathrm{CFG}}(t) = \mathrm{A}(t)\int_0^1 \frac{\mathbb{C}\left(u^{1-t}, u^t\right)du}{u\log(u)} \tag{6}$$

See Genest and Segers (2009, Theorem 3.2) and Gudendorf and Segers (2012, Theorem 1).

In the case of known margins, the results (3) and (4) hold with $\hat{\mathrm{C}}_n$ replaced by C_n and thus $\mathbb{C}_n$ replaced by α_n, as well as process $\mathbb{C}$ replaced by α in (5) and (6). These were already proved in Segers (2007).

The new estimator can be stated for the Pickands dependence function based on Ferreira and Ferreira (2012), and will be denoted FF. Define

$$\eta(t) = U^{1/(1-t)} \vee V^{1/t}$$

with $\eta(0) = U$ and $\eta(1) = V$. By Proposition 3.1 of Ferreira and Ferreira (2012), we have

$$\mathrm{E}\left(\eta(t)\right) = 1 - \frac{1}{1+\mathrm{A}(t)}$$

By an analogous reasoning used above, let

$$\eta_i(t) = U_i^{1/(1-t)} \vee V_i^{1/t}, \quad 0 < t < 1$$

with $\eta_i(0) = U_i$ and $\eta_i(1) = V_i$, $i = 1,\ldots, n$. Thus, in the case of known margins we derive

$$1 - \frac{1}{\left(1+\mathrm{A}_n^{\mathrm{FF}}(t)\right)} = \frac{1}{n}\sum_{i=1}^{n}\eta_i(t)$$

and, for unknown margins,

$$1 - \frac{1}{\left(1+\hat{\mathrm{A}}_n^{\mathrm{FF}}(t)\right)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i(t)$$

where $\hat{\eta}_i(0) = \hat{U}_i$, $\hat{\eta}_i(1) = \hat{V}_i$, and

$$\hat{\eta}_i(t) = \hat{U}_i^{1/(1-t)} \vee \hat{V}_i^{1/t}, \quad 0 < t < 1$$

with $\hat{U}_i$ and $\hat{V}_i$ as defined in (2). Because

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i(0) = \frac{1}{n}\sum_{i=1}^{n}\hat{U}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i(1) = \frac{1}{n}\sum_{i=1}^{n}\hat{V}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{n+1} = \frac{1}{2}$$

the estimator already satisfies the constraints $\mathrm{A}_n^{\mathrm{FF}}(0) = \mathrm{A}_n^{\mathrm{FF}}(1) = 1$. The following statements are direct adaptations of the results above concerning Pickands and CFG estimators.

Proposition 1: For all $t \in [0, 1]$,

$$\sqrt{n}\left(\frac{1}{1+\hat{\mathrm{A}}_n^{\mathrm{FF}}(t)} - \frac{1}{1+\mathrm{A}(t)}\right) = \int_0^1 \mathbb{C}_n\left(u^{1-t}, u^t\right) du \tag{7}$$

Proof: Observe that

$$\frac{1}{1+\hat{\mathrm{A}}_n^{\mathrm{FF}}(t)} = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathbb{I}_{\{\hat{\eta}_i(t)\le u\}} du = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathbb{I}_{\left\{\hat{U}_i \le u^{1-t}, \hat{V}_i \le u^t\right\}} du = \int_0^1 \hat{\mathrm{C}}_n\left(u^{1-t}, u^t\right) du$$

Proposition 2: If A is twice continuously differentiable on (0, 1) such that $\sup_{0<t<1} t(1-t)\mathrm{A}''(t) < \infty$, we have

$$\mathbb{A}_n^{\mathrm{FF}} = \sqrt{n}\left(\hat{\mathrm{A}}_n^{\mathrm{FF}}(t) - \mathrm{A}(t)\right) \xrightarrow{w} \mathbb{A}_n^{\mathrm{FF}}(t) = -\left(\mathrm{A}(t)+1\right)^2 \int_0^1 \mathbb{C}\left(u^{1-t}, u^t\right) du \tag{8}$$

in $\mathcal{C}([0, 1])$ equipped with the topology of uniform convergence.

Proof: Considering $u = \mathrm{e}^{-s}$ in the integral of (7),

$$\sqrt{n}\left(\frac{1}{1+\hat{\mathrm{A}}_n^{\mathrm{FF}}(t)}-\frac{1}{1+\mathrm{A}(t)}\right)=\int_0^\infty \mathbb{C}_n\left(\mathrm{e}^{-s(1-t)},\mathrm{e}^{-st}\right)\mathrm{h}(s)\,ds \tag{9}$$

with $\mathrm{h}(s)=\mathrm{e}^{-s}$. The proof of the convergence of the integral in (9) towards $\int_0^\infty \mathbb{C}\left(\mathrm{e}^{-s(1-t)},\mathrm{e}^{-st}\right)\mathrm{h}(s)\,ds$ runs as the one of Theorem 1 in Gudendorf and Segers (2012). Now the assertion follows by applying the functional delta method (van der Vaart & Wellner, 1996).

For the case of known margins, replace $\hat{\mathrm{C}}_n$ by C_n, $\mathbb{C}_n$ by α_n, and $\mathbb{C}$ by α, respectively, in (7) and (8). See Gudendorf and Segers (2012) and references therein. Furthermore, Propositions 1 and 2 are extensible to the d-variate case for $d > 2$ as stated, respectively, in Lemma 1 and Theorem 1 of Gudendorf and Segers (2012).

Simulations

Consider the most interesting case for practical purposes of unknown margins, where the performance of the new estimator is examined through simulation and compared with the corrected version of CFG and Hall and Tajvidi estimators. Specifically, 1000 random samples of size $n = 100$, and of $n = 1000$ were generated for each of the following models: logistic, asymmetric logistic, Hüsler-Reiss, negative logistic, asymmetric negative logistic, bilogistic, negative bilogistic, Dirichlet, and asymmetric mixed. A description of the latter can be found in Beirlant et al. (2004).

The empirical mean integrated squared error, $\mathrm{MISE}=\mathrm{E}\left(\int_0^1\left(\hat{\mathrm{A}}_n(t)-\mathrm{A}(t)\right)^2 dt\right)$, was computed for each estimator and the obtained values are reported in Tables 1-3 (the numbers in brackets correspond to standard errors). The values of the parameters of each model were chosen in order to have the TDC ($\lambda = 2(1 - \mathrm{A}(0.5))$) approximately 0.5 and the boundary cases 0 and 1, corresponding to Tables 1, 2, and 3, respectively. In the unit bound case in Table 3, i.e., $\lambda \approx 1$, the considered asymmetric versions coincide with the respective symmetric models and thus omitted. Also, in the asymmetric mixed model, the largest value achieved by λ correspond to 0.5 already reported in Table 1. Observe that the unit TDC scenario presents the smallest errors. Note the FF estimator has an overall good performance, particularly in the boundary cases of asymptotic tail independence ($\lambda \approx 0$) and $\lambda \approx 1$ (see Tables 2 and 3).

Table 1. Empirical MISE values obtained for estimators CFG, HT and FF of the Pickands dependence function where the considered parameters for each model are such that $\lambda \approx 0.5$

n = 1000		CFG		HT		FF
Log	4.070×10^{-5}	(3.011×10^{-6})	5.607×10^{-5}	(5.607×10^{-6})	4.569×10^{-5}	(4.569×10^{-6})
Alog	8.383×10^{-4}	(6.200×10^{-5})	8.403×10^{-4}	(6.199×10^{-5})	8.496×10^{-4}	(6.268×10^{-5})
HR	3.587×10^{-5}	(3.046×10^{-6})	4.840×10^{-5}	(4.170×10^{-6})	3.947×10^{-5}	(3.364×10^{-6})
Neglog	4.181×10^{-5}	(3.306×10^{-6})	5.560×10^{-5}	(4.444×10^{-6})	4.609×10^{-5}	(3.669×10^{-6})
Aneglog	6.809×10^{-5}	(3.952×10^{-6})	8.318×10^{-5}	(4.819×10^{-6})	6.858×10^{-5}	(3.995×10^{-6})
Bilog	5.032×10^{-4}	(3.897×10^{-5})	5.221×10^{-4}	(3.942×10^{-5})	5.115×10^{-4}	(3.948×10^{-5})
Negbilog	1.063×10^{-4}	(6.854×10^{-6})	1.200×10^{-4}	(7.558×10^{-6})	1.123×10^{-4}	(7.204×10^{-6})
Dir	4.114×10^{-4}	(3.114×10^{-5})	4.342×10^{-4}	(3.205×10^{-5})	4.191×10^{-4}	(3.150×10^{-5})
Amix	4.156×10^{-5}	(3.063×10^{-6})	5.621×10^{-5}	(4.319×10^{-6})	4.604×10^{-5}	(3.401×10^{-6})

n = 100		CFG		HT		FF
Log	2.890×10^{-4}	(2.861×10^{-6})	4.181×10^{-4}	(4.140×10^{-6})	3.656×10^{-4}	(3.323×10^{-6})
Alog	1.289×10^{-3}	(7.866×10^{-5})	1.436×10^{-3}	(8.335×10^{-5})	1.403×10^{-3}	(8.386×10^{-5})
HR	3.544×10^{-4}	(3.035×10^{-5})	4.595×10^{-4}	(4.011×10^{-5})	4.043×10^{-4}	(3.385×10^{-5})
Neglog	3.948×10^{-4}	(3.246×10^{-5})	5.368×10^{-4}	(4.482×10^{-5})	4.584×10^{-4}	(3.713×10^{-5})
Aneglog	6.150×10^{-4}	(3.735×10^{-5})	7.542×10^{-4}	(4.435×10^{-5})	6.787×10^{-4}	(4.027×10^{-5})
Bilog	8.055×10^{-4}	(5.198×10^{-5})	9.542×10^{-4}	(6.003×10^{-5})	8.872×10^{-4}	(5.600×10^{-5})
Negbilog	4.231×10^{-4}	(3.147×10^{-5})	5.505×10^{-4}	(4.182×10^{-5})	4.786×10^{-4}	(3.489×10^{-5})
Dir	7.399×10^{-4}	(4.869×10^{-5})	8.956×10^{-4}	(5.916×10^{-5})	8.117×10^{-4}	(5.214×10^{-5})
Amix	4.249×10^{-4}	(3.462×10^{-5})	5.617×10^{-4}	(4.730×10^{-5})	4.748×10^{-4}	(3.752×10^{-5})

Note: Numbers in brackets correspond to standard errors

Table 2. Empirical MISE values obtained for estimators CFG, HT and FF of the Pickands dependence function, in the case of asymptotic tail independence ($\lambda \approx 0$)

n = 1000		CFG		HT		FF
Log	1.020×10^{-4}	(5.090×10^{-6})	1.997×10^{-4}	(1.017×10^{-5})	7.133×10^{-5}	(3.616×10^{-6})
Alog	9.932×10^{-5}	(4.885×10^{-6})	2.103×10^{-4}	(1.042×10^{-5})	6.230×10^{-5}	(3.007×10^{-6})
HR	1.054×10^{-4}	(5.203×10^{-6})	2.212×10^{-4}	(1.068×10^{-5})	7.121×10^{-5}	(3.499×10^{-6})
Neglog	1.021×10^{-4}	(5.161×10^{-6})	2.052×10^{-4}	(1.065×10^{-5})	6.792×10^{-5}	(3.502×10^{-6})
Aneglog	1.032×10^{-4}	(5.171×10^{-6})	2.101×10^{-4}	(1.081×10^{-5})	6.890×10^{-5}	(3.468×10^{-6})
Bilog	1.025×10^{-4}	(5.413×10^{-5})	2.093×10^{-4}	(1.145×10^{-5})	7.438×10^{-5}	(4.023×10^{-6})
Negbilog	1.042×10^{-4}	(5.279×10^{-6})	2.142×10^{-4}	(1.123×10^{-5})	7.067×10^{-5}	(3.565×10^{-6})
Dir	1.022×10^{-4}	(5.162×10^{-5})	2.072×10^{-4}	(1.060×10^{-5})	7.737×10^{-5}	(4.080×10^{-6})
Amix	1.054×10^{-4}	(5.248×10^{-6})	2.100×10^{-4}	(1.054×10^{-5})	7.307×10^{-5}	(3.698×10^{-6})

Note: Numbers in brackets correspond to standard errors

Table 2, continued.

n = 100		CFG		HT		FF
Log	1.404×10^{-3}	(6.483×10^{-5})	2.232×10^{-3}	(1.120×10^{-4})	9.676×10^{-4}	(4.309×10^{-5})
Alog	1.349×10^{-3}	(6.389×10^{-5})	2.165×10^{-3}	(1.107×10^{-4})	9.250×10^{-4}	(4.308×10^{-5})
HR	1.350×10^{-3}	(6.161×10^{-5})	2.121×10^{-3}	(1.096×10^{-4})	9.128×10^{-4}	(3.889×10^{-5})
Neglog	1.344×10^{-3}	(6.274×10^{-5})	2.181×10^{-3}	(1.128×10^{-4})	8.938×10^{-4}	(3.966×10^{-5})
Aneglog	1.441×10^{-3}	(6.655×10^{-5})	2.141×10^{-3}	(1.064×10^{-4})	1.001×10^{-4}	(4.488×10^{-5})
Bilog	1.339×10^{-3}	(6.351×10^{-5})	2.123×10^{-3}	(1.052×10^{-4})	9.496×10^{-4}	(4.462×10^{-5})
Negbilog	1.236×10^{-3}	(5.570×10^{-5})	1.989×10^{-3}	(1.035×10^{-4})	8.316×10^{-4}	(3.542×10^{-5})
Dir	1.345×10^{-3}	(6.343×10^{-5})	2.087×10^{-3}	(1.047×10^{-4})	9.509×10^{-4}	(4.345×10^{-5})
Amix	1.409×10^{-3}	(6.608×10^{-5})	2.190×10^{-3}	(1.139×10^{-4})	9.676×10^{-4}	(4.348×10^{-5})

Note: Numbers in brackets correspond to standard errors

Table 3. Empirical MISE values obtained for estimators CFG, HT and FF of the Pickands dependence function, in the case $\lambda \approx 1$

n = 100		CFG		HT		FF
Log	3.874×10^{-9}	(2.394×10^{-9})	3.539×10^{-9}	(2.262×10^{-9})	6.118×10^{-10}	(3.926×10^{-10})
HR	4.930×10^{-10}	(2.935×10^{-9})	4.413×10^{-9}	(2.768×10^{-9})	5.571×10^{-10}	(3.625×10^{-10})
Neglog	4.001×10^{-9}	(2.451×10^{-9})	3.709×10^{-9}	(2.378×10^{-9})	5.826×10^{-10}	(3.753×10^{-10})
Bilog	3.913×10^{-9}	(2.400×10^{-10})	3.610×10^{-9}	(2.312×10^{-9})	6.220×10^{-10}	(4.000×10^{-10})
Negbilog	4.131×10^{-9}	(2.464×10^{-9})	3.517×10^{-9}	(2.276×10^{-9})	5.985×10^{-10}	(2.869×10^{-10})
Dir	2.890×10^{-8}	(2.612×10^{-8})	2.154×10^{-7}	(3.900×10^{-8})	8.186×10^{-8}	(2.721×10^{-8})

n = 100		CFG		HT		FF
Log	1.530×10^{-7}	(7.468×10^{-8})	1.352×10^{-7}	(7.366×10^{-8})	1.342×10^{-8}	(1.062×10^{-8})
HR	1.872×10^{-7}	(9.113×10^{-8})	1.627×10^{-7}	(8.760×10^{-8})	1.903×10^{-8}	(1.487×10^{-8})
Neglog	1.492×10^{-7}	(7.352×10^{-8})	1.348×10^{-7}	(7.360×10^{-8})	1.279×10^{-8}	(1.018×10^{-8})
Bilog	1.516×10^{-7}	(7.519×10^{-8})	1.342×10^{-7}	(7.256×10^{-8})	1.606×10^{-8}	(1.035×10^{-8})
Negbilog	1.519×10^{-7}	(7.513×10^{-8})	1.361×10^{-7}	(7.457×10^{-8})	1.265×10^{-8}	(1.003×10^{-8})
Dir	2.074×10^{-6}	(5.517×10^{-7})	2.033×10^{-6}	(6.234×10^{-7})	1.250×10^{-6}	(4.285×10^{-7})

Note: Numbers in brackets correspond to standard errors

Conclusion

A new estimator for the Pickands dependence function, an important map in generating extreme value copulas, was presented. It was found via simulation that it may be used as an alternative to the well-known CFG estimator, especially in the limiting situation of asymptotic tail independence. Thus, it may have a promising performance in testing independence, a crucial issue in statistics of extremes.

Acknowledgements

The author wishes to thank the reviewers for their careful reading and relevant comments that have improved this work. This research was financed by Portuguese Funds through FCT – Fundação para a Ciência e a Tecnologia within the Projects UID/MAT/00013/2013, UID/MAT/00006/2013, and UID/Multi/04621/2013.

References

Abdous, B., & Ghoudi, K. (2005). Non-parametric estimators of multivariate extreme dependence functions. *Journal of Nonparametric Statistics, 17*(8), 915-935. doi: 10.1080/10485250500336379

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2004). *Statistics of extremes: Theory and applications*. Chichester, UK: Wiley. doi: 10.1002/0470012382

Capéraà, P., Fougères, A.-L., & Genest, C. (1997). A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika, 84*(3), 567-577. doi: 10.1093/biomet/84.3.567

Cebrian, A., Denuit, M., & Lambert, P. (2003). Analysis of bivariate tail dependence using extreme values copulas: An application to the SOA medical large claims database. *Belgian Actuarial Bulletin, 3*(1), 33-41.

Coles, S. G., Heffernan, J., & Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes, 2*(4), 339-365. doi: 10.1023/a:1009963131610

Deheuvels, P. (1984). Probabilistic aspects of multivariate extremes. In J. T. de Oliveira (Ed.), *Statistical extremes and applications* (pp. 117-130). Boston, MA: D. Reidel Pub. Co. doi: 10.1007/978-94-017-3069-3_9

Deheuvels, P. (1991). On the limiting behavior of the Pickands estimator for bivariate extreme-value distributions. *Statistics & Probability Letters, 12*(5), 429-439. doi: 10.1016/0167-7152(91)90032-m

Ferreira, H. & Ferreira, M. (2012). On extremal dependence of block vectors. *Kybernetika, 48*(5), 988-1006. Retrieved from http://www.kybernetika.cz/content/2012/5/988

Frees, E. W., & Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal, 2*(1), 1-25. doi: 10.1080/10920277.1998.10595667

Genest, C., & Segers, J. (2009). Rank-based inference for bivariate extreme-value copulas. *The Annals of Statistics, 37*(5B), 2990-3022. doi: 10.1214/08-aos672

Ghoudi, K., Khoudraji, A., & Rivest, L.-P. (1998). Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. *The Canadian Journal of Statistics, 26*(1), 187-197. doi: 10.2307/3315683

Gudendorf, G., & Segers, J. (2011). Nonparametric estimation of an extreme-value copula in arbitrary dimensions. *Journal of Multivariate Analysis, 102*(1), 37-47. doi: 10.1016/j.jmva.2010.07.011

Gudendorf, G., & Segers, J. (2012). Nonparametric estimation of multivariate extreme-value copulas. *Journal of Statistical Planning and Inference, 142*(12), 3073-3085. doi: 10.1016/j.jspi.2012.05.007

Hall, P., & Tajvidi, N. (2000). Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli, 6*(5), 835-844. doi: 10.2307/3318758

Hsing, T. (1989). Extreme value theory for multivariate stationary sequences. *Journal of Multivariate Analysis, 29*(2), 274-291. doi: 10.1016/0047-259x(89)90028-6

Jiménez, J. R., Villa-Diharce, E., & Flores, M. (2001). Nonparametric estimation of the dependence function in bivariate extreme value distributions. *Journal of Multivariate Analysis, 76*(2), 159-191. doi: 10.1006/jmva.2000.1931

McNeil, A. J., Frey, R., & Embrechts, P. (2005). *Quantitative risk management: Concepts, techniques and tools*. Princeton, NJ: Princeton University Press.

Pickands, J. (1981). Multivariate extreme value distributions. In *Proceedings of the 43rd Session of the International Statistical Institute* (Vol. 2, pp. 859-878). Buenos Aires, Brazil: International Statistical Institute.

Salvadori, G., De Michele, C., Kottegoda, N. T., & Rosso, R. (2007). *Extremes in nature: An approach using copulas*. Dordrecht, Netherlands: Springer. doi: 10.1007/1-4020-4415-1

Segers, J. (2007). Nonparametric inference for bivariate extreme-value copulas. In M. Ahsanullah & S. N. U. A. Kirmani (Eds.), *Topics in extreme values* (pp. 185-207). New York, NY: Nova Science Publishers.

Sibuya, M. (1960). Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics, 11*(2), 195-210. doi: 10.1007/bf01682329

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris, 8*, 229-231.

Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika, 75*(3), 397-415. doi: 10.1093/biomet/75.3.397

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. New York, NY: Springer. doi: 10.1007/978-1-4757-2545-2

Zhang, D., Wells, M. T., & Peng, L. (2008). Nonparametric estimation of the dependence function for a multivariate extreme value distribution. *Journal of Multivariate Analysis, 99*(4), 577-588. doi: 10.1016/j.jmva.2006.09.011

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 364-387.
doi: 10.22237/jmasm/1493598060

ISSN 1538 − 9472

A New Estimator based on Auxiliary Information through Quantitative Randomized Response Techniques

Nilgün Özgül
Hacettepe University
Ankara, Turkey

Hülya Çıngı
Hacettepe University
Ankara, Turkey

An exponential-type estimator is developed for the population mean of the sensitive study variable based on various Randomized Response Techniques (RRT) using a non-sensitive auxiliary variable. The mean squared error (MSE) of the proposed estimator is derived for generalized RRT models. The proposed estimator is compared with competitors in a simulation study and an application. The proposed estimator is found to be more efficient using a non-sensitive auxiliary variable.

Keywords: Randomized response techniques, sensitive question, auxiliary variable, exponential estimator, efficiency

Introduction

In surveys on sensitive topics, estimation of the population mean with a direct questioning technique may cause respondents to refuse answering or to give untruthful answers on purpose. Respondents may encounter questions about drug use, illegal income, political views, abortion, homosexual activities, and AIDS in some social, medical, and epidemiological questionnaires. On these surveys, respondents do not feel comfortable and they may choose not to answer or may intentionally provide false answers. This can bring about significant bias in the estimation of population parameters.

Random response techniques (RRT) are used to reduce nonrespondent's rates and biased responses to sensitive questions. Warner (1965) introduced the randomization technique for the proportion of a population characterized by a sensitive variable, which was followed by studies where the response to a sensitive question results in a quantitative variable. Quantitative RRT are used to

Dr. Özgül one is a Professor in the Department of Statistics. Email them at: nilgunozgul@yahoo.com.

estimate the mean value of some behavior in a population. For example, the sensitive study variable may be the total number of abortions a woman has had or the average weekly alcohol consumption or annual earnings of people. These RRT are sub-classified as either additive or multiplicative techniques.

In additive RRT, respondents are asked to scramble their responses using a randomization device such as a deck of cards. Each of the cards in the deck has a number. The numbers in the deck follow a known probability distribution, such as Normal, Chi-square, Uniform, Poisson, Binomial, Weibull, etc. The respondent is asked to add the real response to the number listed on card picked, and then report only the sum to the interviewer. Multiplicative RRT are similar to additive RRT. Again, a deck of cards with known probability distribution is used, but now when the respondents scramble their responses, they are asked to report the product of the real response and the number listed on the selected card. The interviewer cannot see the card, but records the reported number. RRT can also be categorized by how the respondents are instructed to randomize. If all respondents are asked to randomize their response, the model is characterized as a full randomization RRT model. If some of the respondents are instructed to randomize their response, the model is characterized as a "partial RRT model" (Özgül, 2013).

Thornton and Gupta (2004) extended Warner's (1971) approach by using partial additive models for estimating the mean of sensitive quantitative variables in RRT. The multiplicative model was later investigated in depth by Eichhorn and Hayre (1983), who referred to it as the scrambled responses method. Similarly, Bar-Lev, Bobovitch, and Boukai (2004) proposed a method which uses a partial model that generalizes Eichhorn and Hayre's results and yields an estimate which, under mild conditions, has a uniformly smaller variance. Further developments focused on the use of auxiliary variables to improve the precision. Diana and Perri (2011), Sousa, Shabbir, Real, and Gupta (2010), and Gupta, Shabbir, Sousa, and Real (2012) suggested mean estimators using the auxiliary variable for estimating of the quantitative sensitive variable in RRT. Bahl and Tuteja (1991), Shabbir and Gupta (2011), Grover and Kaur (2014), and Özgül and Cingi (2014) studied exponential-type estimators to obtain more efficient estimates for various sampling methods. In the current study, an exponential-type estimator of the mean of a sensitive variable is proposed using a non-sensitive auxiliary variable for generalized partial quantitative RRT.

Various Estimators Based on Auxiliary Information through Quantitative RRT

Diana and Perri (2011) introduced a general mechanism to scramble responses and proposed a class of regression estimators for the mean of a sensitive variable using a non-sensitive auxiliary variable. To estimate μ_y, a sample of individuals is selected from the population and each respondent is asked to perform a Bernoulli trial with a probability of success P. If this is successful, the respondent then gives the true values of both Y and X. In the case of failure, the respondent gives their answers by using the values given in S and R, which are the various randomized designs for the variables Y and X, respectively. The interviewer does not know the outcome of the Bernoulli experiment. Then, the distribution of the responses is given in (1) as

$$(Z,U) \rightarrow \begin{cases} (Y,X) & \text{with probability } P \\ (S,R) & \text{with probability } (1-P) \end{cases} \tag{1}$$

where Y is the sensitive variable of interest with unknown mean μ_y and unknown variance S_y^2, X is the non-sensitive variable with known mean μ_x and known variance S_x^2, Z is the reported response for the sensitive variable Y, and U is the reported response for the first non-sensitive variable X. In S and R, the respondents answer the questions using the additive or multiplicative technique. For the additive technique, each respondent is requested to draw a value from the distribution of the scrambling variable, add it to the real response, and report back to the interviewer. For the multiplicative model, the respondent responds with the product of the drawn value and their true response. The scrambling variables are defined as W and T which have pre-assigned distributions such as Normal, Chi-square, Uniform, Poisson, Binomial, Weibull, etc. W is the scrambling variable with known true mean μ_w and known variance S_w^2 in S and T is the scrambling variable with known true mean μ_t and variance S_t^2 in R (Özgül, 2013).

Under the generic scheme given in (1), the following class of estimators based on a SRSWR sample $\{(z_1, u_1), (z_2, u_2), \ldots, (z_n, u_n)\}$ of n responses is

$$\hat{\mu}_{\mathrm{DP}} = \frac{\bar{z} + b(\mu_u - \bar{u}) - c}{h}, \quad (h \neq 0) \tag{2}$$

where b is a suitably selected real constant and

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad \bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

are the sample means of the reported responses for the sensitive variable and the non-sensitive auxiliary variable, respectively. Here, c and h depend exclusively on the scrambling design.

The variance of $\hat{\mu}_{\mathrm{DP}}$ is

$$\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right) = \frac{1}{nh^2}\left(S_z^2 - 2BS_{zu} + B^2 S_u^2\right) \tag{3}$$

where

$$S_z^2 = \frac{\sum_{i=1}^{N}\left(z_i - \mu_z\right)^2}{N-1}, \quad S_u^2 = \frac{\sum_{i=1}^{N}\left(u_i - \mu_z\right)^2}{N-1}$$

are the population variances of z and u, respectively,

$$S_{zu} = \frac{\sum_{i=1}^{N}\left(z_i - \mu_z\right)\left(u_i - \mu_u\right)}{N-1}$$

is the population covariance between z and u, $B = S_{zu}/S_u^2$ is the population regression coefficient between z and u, and

$$\mu_z = \frac{1}{N}\sum_{i=1}^{N} z_i, \quad \mu_u = \frac{1}{N}\sum_{i=1}^{N} u_i$$

are the population means of z and u, respectively. The minimum variance of $\hat{\mu}_{\mathrm{DP}}$ is

$$\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min} = \frac{S_z^2}{nh^2}\left(1 - \rho_{zu}^2\right) \tag{4}$$

where $\rho_{zu} = S_{zu} / \left(S_z S_u \right)$ is the population correlation coefficient between z and u.

Sousa et al. (2010) proposed a ratio estimator for the mean of a sensitive variable using a non-sensitive auxiliary variable. The respondent is asked to provide true responses for X. The Sousa et al. estimator is

$$\hat{\mu}_{\mathrm{SR}} = \overline{z} \left(\frac{\mu_x}{\overline{x}} \right) \tag{5}$$

where $\bar{z}$ is the sample mean of the reported responses for the sensitive variable ($Z = Y + W$),

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i$$

is the known population mean of non-sensitive auxiliary variable, and

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is the sample mean of non-sensitive auxiliary variable. The Bias and MSE of $\hat{\mu}_{\mathrm{SR}}$, under first order of the approximation, is

$$\mathrm{Bias}\left(\hat{\mu}_{SR} \right) \cong \lambda \mu_z \left(C_x^2 - C_{zx} \right) \tag{6}$$

$$\mathrm{MSE}\left(\hat{\mu}_{SR} \right) = \lambda \mu_z^2 \left[C_z^2 + C_x^2 - 2 \rho_{xz} C_x C_z \right] \tag{7}$$

where

$$\lambda = \frac{1}{n} - \frac{1}{N}$$

and $C_z = S_z/\mu_z$ and $C_x = S_x/\mu_x$ are the coefficients of variation of Z and X, respectively.

Gupta et al. (2012) proposed regression-cum-ratio estimator using a non-sensitive auxiliary variable.

$$\hat{\mu}_{\text{GRR}} = \left[b_1 \bar{z} + b_2 \left(\mu_x - \bar{x} \right) \right] \left(\frac{\mu_x}{\bar{x}} \right) \tag{8}$$

where $\bar{z}$, μx, and $\bar{x}$ are defined as above for (5), and b_1 and b_2 are constants. The Bias and minimum MSE of $\hat{\mu}_{\text{GRR}}$, under first order of the approximation, is

$$\text{Bias}\left(\hat{\mu}_{GRR} \right) \cong \left(b_1 - 1 \right) \mu_z + b_1 \lambda \mu_z \left\{ C_x^2 - C_{zx} \right\} + b_2 \lambda \mu_x C_x^2 \tag{9}$$

$$\text{MSE}\left(\hat{\mu}_{GRR} \right)_{\min} \cong \frac{\lambda \mu_z^2 C_z^2 \left(1 - \rho_{zx}^2 \right) \left(1 - \lambda C_x^2 \right)}{\lambda C_z^2 \left(1 - \rho_{zx}^2 \right) + \left(1 - \lambda C_x^2 \right)} \tag{10}$$

Suggested Exponential-Type Estimator Based on Auxiliary Information through Quantitative RRT

Applying the general formulation of Diana and Perri (2011) and following Grover and Kaur (2014), an exponential-type estimator for the mean of a sensitive variable is proposed using a non-sensitive auxiliary variable in RRT. Consider the following improved exponential estimator based on a SRSWOR sample $\{(z_1, u_1), (z_2, u_2), \ldots, (z_n, u_n)\}$ of n responses:

$$\hat{\mu}_{\text{NH(exp)}} = \frac{\left[\left\{ b_1 \bar{z} + b_2 \left(\mu_u - \bar{u} \right) \right\} \exp\left(\frac{\alpha \left(\mu_u - \bar{u} \right)}{\alpha \left(\mu_u + \bar{u} \right) + \beta} \right) \right] - c}{h}, \quad \left(h \neq 0 \right) \tag{11}$$

where b_1 and b_2 are suitably selected real constant and α and β are already assumed to be either any known constants or functions of any known population parameters of the auxiliary variable, such as standard deviation (σ_x), coefficient of variation (C_x), coefficient of skewness $\{\beta_1(x)\}$, coefficient of kurtosis $\{\beta_2(x)\}$, coefficient of correlation (ρ_{yx}) (Cingi & Kadilar, 2009). Here, c and h depend exclusively on the scrambling design.

To obtain the MSE equation for the proposed estimator, we define following relative error terms

$$e_0 = \frac{\left(\bar{z} - \mu_z \right)}{\mu_z}, e_1 = \frac{\left(\bar{u} - \mu_u \right)}{\mu_u}$$

such that

$$\mathrm{E}(e_0)=\mathrm{E}(e_1)=0;\ \mathrm{E}(e_0^2)=\lambda C_z^2,\ \mathrm{E}(e_1^2)=\lambda C_u^2,\ \mathrm{E}(e_0e_1)=\lambda\rho_{zu}C_zC_u$$

where

$$\rho_{zu}=\frac{S_{zu}}{S_zS_u},\ C_u^2=\frac{S_u^2}{\mu_u^2},\ C_u=\frac{S_u}{\mu_u},\ C_z=\frac{S_z}{\mu_z},$$
$$S_z^2=\frac{\sum_{i=1}^{N}(z_i-\mu_z)^2}{N-1},\ S_u^2=\frac{\sum_{i=1}^{N}(u_i-\mu_z)^2}{N-1}$$

Expressing (11) in terms of the *e*'s:

$$\begin{aligned}\hat{\mu}_{\mathrm{NH(exp)}}&=\frac{\left\{b_1\bar{Z}(1+e_0)-b_2\mu_ue_1\right\}\exp\left\{-\gamma e_1(1+\gamma e_1)^{-1}\right\}-c}{h}\\&=\frac{\left\{b_1\bar{Z}(1+e_0)-b_2\mu_ue_1\right\}\left\{1-\gamma e_1+\tfrac{3}{2}\gamma^2e_1^2+\ldots\right\}-c}{h}\end{aligned}\qquad(12)$$

where

$$\gamma=\frac{\alpha\mu_u}{2(\alpha\mu_u+\beta)}$$

Assuming $|e_1|<1$, expanding the right hand side of (10), and retaining terms up to the second degree of the *e*'s we have

$$\begin{aligned}\hat{\mu}_{\mathrm{NH(exp)}}-\mu_z\cong\mu_z\left[(b_1-1)+b_1e_0-b_1\gamma(e_1+e_0e_1)+\tfrac{3}{2}b_1\gamma^2e_1^2\right]\\+b_2\mu_u\left(-e_1-\gamma e_1^2\right)\end{aligned}\qquad(13)$$

Taking the expectation both sides of (13), the Bias Equation of $\hat{\mu}_{\mathrm{NH(exp)}}$ is obtained to the first degree of approximation as

$$\text{Bias}\left(\hat{\mu}_{\text{NH(exp)}}\right) \cong \frac{\mu_z\left\{\left(b_1-1\right)+\lambda b_1\gamma C_u\left(\frac{3}{2}\gamma C_u-\rho_{zu}C_z\right)\right\}+\lambda b_2\mu_u\gamma C_u^2-c}{h} \quad (14)$$

Squaring both sides of (13), retaining terms of the *e*'s up to the second degree and taking the expectation, we get the MSE Equation of $\hat{\mu}_{\text{NH(exp)}}$ to the first degree of approximation as

$$\begin{aligned}\text{MSE}\left(\hat{\mu}_{\text{NH(exp)}}\right) = \frac{\mu_z^2}{h^2}\Big[\left(b_1-1\right)^2+\lambda b_1^2 A-\lambda\gamma b_1\left\{D+\left(\gamma C_u^2-\rho_{zu}C_zC_u\right)\right\}\Big] \\ +\lambda b_2^2\mu_u^2C_u^2+2\lambda b_2\mu_z\mu_u\left[b_1D-\gamma C_u^2\right]\end{aligned} \quad (15)$$

where $A=C_z^2+4\gamma^2C_u^2-4\gamma\rho_{zu}C_zC_u$, $D=2\gamma C_u^2-\rho_{zu}C_zC_u$, and C_u is the coefficient of variation of *u*.

To minimize $\text{MSE}\left(\hat{\mu}_{\text{NH(exp)}}\right)$, consider the following normal equations:

$$\frac{\partial MSE\left(\hat{\mu}_{\text{NH(exp)}}\right)}{\partial b_i}=0, \quad i=1,2$$

On solving these two normal equations simultaneously, the optimum values of b_1 and b_2 are, respectively,

$$\begin{aligned}b_{1(\text{opt})} &= \frac{-C_u^2\left[2-\lambda\gamma D+\lambda\gamma\left(\gamma C_u^2-\rho_{zu}C_zC_u\right)\right]}{2\left[\lambda D^2-\left(1+\lambda A\right)C_u^2\right]} \\ b_{2(\text{opt})} &= \frac{\mu_z\left[D\left\{2+\lambda\gamma D+\lambda\gamma\left(\gamma C_u^2-\rho_{zu}C_zC_u\right)\right\}-2\gamma C_u^2\left(1+\lambda A\right)\right]}{2\mu_u\left[\lambda D^2-\left(1+\lambda A\right)C_u^2\right]}\end{aligned} \quad (16)$$

On substituting the optimum values of b_1 and b_2 from (15) into (14), the minimum MSE of the proposed estimator $\hat{\mu}_{\text{NH(exp)}}$, up to first order of approximation, is given by

$$\text{MinMSE}\left(\hat{\mu}_{\text{NH(exp)}}\right)=\frac{1}{h^2}\left[\begin{array}{l}\dfrac{\lambda\mu_z^2C_z^2\left(1-\rho_{zu}^2\right)}{1+\lambda C_z^2\left(1-\rho_{zu}^2\right)}\\ \qquad -\dfrac{\lambda^2\mu_z^2\gamma^2C_u^2\left\{4C_z^2\left(1-\rho_{zu}^2\right)+\gamma^2C_u^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zu}^2\right)\right\}}\end{array}\right] \quad (17)$$

The expressions of c, h, and the MSE and mean equations change depending on the specified models. Two additive models and two multiplicative models are specified. In the first model, M_1, the additive technique is applied for the sensitive variable while the direct technique is utilized for the non-sensitive auxiliary variable: $\{Z = PY + (1 - P)(Y + W),\ U = X\}$. In the second model, M_2, the multiplicative model is applied for the sensitive variable while the direct technique is utilized for the non-sensitive auxiliary variable: $\{Z = PY + (1 - P)(YW),\ U = X\}$. In the third model, M_3, the additive model is applied for both the sensitive variable and the non-sensitive auxiliary variable: $\{Z = PY + (1 - P)(Y + W),\ U = PX + (1 - P)(X + T)\}$. In the fourth model, M_4, the multiplicative model is applied for both the sensitive variable and the non-sensitive auxiliary variable: $\{Z = PY + (1 - P)(YW),\ U = PX + (1 - P)(XT)\}$. In some surveys dealing with sensitive topics, the auxiliary variable that researchers determine to be non-sensitive may be sensitive for respondents. Therefore, in the third model M_3 and fourth model M_4, randomized devices are also used for the auxiliary variable. Mean, variance, and correlation equations, which will be used in MSE equation in (17), are presented in Appendix A according to these four models (Özgül, 2013).

Efficiency Comparisons

A comparison of the proposed estimator with the Diana and Perri (2011) estimator $\hat{\mu}_{\text{DP}}$, the Sousa et al. (2010) estimator $\hat{\mu}_{\text{SR}}$, and the Gupta et al. (2012) estimator $\hat{\mu}_{\text{GRR}}$ is now considered. To compare the efficiencies of the various existing estimators with the proposed estimator, we compare their MSE under the model 1 M_1, in which the respondent is asked to provide true responses for X. The MSEs of estimators under that model with SRSWOR are given below:

$$\text{Var}\left(\hat{\mu}_{\text{DP}}\right)_{\min}=\lambda S_z^2\left(1-\rho_{zx}^2\right) \quad (18)$$

$$\mathrm{MSE}\left(\hat{\mu}_{\mathrm{SR}}\right)=\lambda\mu_z^2\left(C_z^2+C_x^2-2\rho_{xz}C_xC_z\right) \tag{19}$$

$$\mathrm{MSE}\left(\hat{\mu}_{\mathrm{GRR}}\right)_{\min}\cong\frac{\lambda\mu_z^2C_z^2\left(1-\rho_{zx}^2\right)\left(1-\lambda C_x^2\right)}{\lambda C_z^2\left(1-\rho_{zx}^2\right)+\left(1-\lambda C_x^2\right)} \tag{20}$$

$$\begin{aligned}\mathrm{MinMSE}\left(\hat{\mu}_{\mathrm{NH(exp)}}\right)&=\left[\begin{array}{c}\dfrac{\lambda\mu_z^2C_z^2\left(1-\rho_{zx}^2\right)}{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)}\\ -\dfrac{\lambda^2\mu_z^2\gamma^2C_x^2\left\{4C_z^2\left(1-\rho_{zx}^2\right)+\gamma^2C_x^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right\}}\end{array}\right]\\&=\lambda\mu_z^2C_z^2\left(1-\rho_{zx}^2\right)-\frac{\lambda^2\mu_z^2C_z^4\left(1-\rho_{zx}^2\right)^2}{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)}-\frac{\lambda^2\mu_z^2\gamma^2C_x^2\left\{4C_z^2\left(1-\rho_{zx}^2\right)+\gamma^2C_x^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right\}}\\&=\lambda\mu_z^2C_z^2\left(1-\rho_{zx}^2\right)-\frac{\dfrac{Var\left(\hat{\mu}_{DP}\right)_{\min}^2}{\mu_z^2}}{1+\dfrac{Var\left(\hat{\mu}_{DP}\right)_{\min}}{\mu_z^2}}-\frac{\lambda^2\mu_z^2\gamma^2C_x^2\left\{4C_z^2\left(1-\rho_{zx}^2\right)+\gamma^2C_x^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right\}}\end{aligned} \tag{21}$$

From (18) and (21),

$$\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min} - \mathrm{MinMSE}\left(\hat{\mu}_{\mathrm{NH(exp)}}\right) = \frac{\dfrac{Var\left(\hat{\mu}_{DP}\right)^2_{\min}}{\mu_z^2}}{1+\dfrac{Var\left(\hat{\mu}_{DP}\right)_{\min}}{\mu_z^2}} + \frac{\lambda^2\mu_z^2\gamma^2C_x^2\left\{4C_z^2\left(1-\rho_{zx}^2\right)+\gamma^2C_x^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right\}}$$

and so $\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min} - \mathrm{MinMSE}\left(\hat{\mu}_{\mathrm{NH(exp)}}\right) > 0$ always.

From (19) and (21),

$$\begin{aligned}\mathrm{MSE}\left(\hat{\mu}_{\mathrm{SR}}\right) - \mathrm{MinMSE}\left(\hat{\mu}_{\mathrm{NH(exp)}}\right) &= \lambda\mu_z^2\left[C_z^2 + C_x^2 - 2\rho_{xz}C_xC_z\right] - \lambda\mu_z^2C_z^2\left(1-\rho_{zx}^2\right) \\ &\quad + \frac{\dfrac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)^2_{\min}}{\mu_z^2}}{1+\dfrac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}}{\mu_z^2}} + \frac{\lambda^2\mu_z^2\gamma^2C_x^2\left\{4C_z^2\left(1-\rho_{zx}^2\right)+\gamma^2C_x^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right\}} \\ &= \lambda\mu_z^2\left(C_x - 2\rho_{xz}C_z\right)^2 + \frac{\dfrac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)^2_{\min}}{\mu_z^2}}{1+\dfrac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}}{\mu_z^2}} + \frac{\lambda^2\mu_z^2\gamma^2C_x^2\left\{4C_z^2\left(1-\rho_{zx}^2\right)+\gamma^2C_x^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right\}}\end{aligned}$$

and so $\mathrm{MSE}\left(\hat{\mu}_{\mathrm{SR}}\right) - \mathrm{MinMSE}\left(\hat{\mu}_{\mathrm{NH(exp)}}\right) > 0$ always.

From (20) and (21),

$$\begin{aligned}
&\mathrm{MSE}\left(\hat{\mu}_{\mathrm{GRR}}\right)_{\min} - \mathrm{MinMSE}\left(\hat{\mu}_{\mathrm{NH(exp)}}\right) \\
&= \frac{\lambda\mu_z^2 C_z^2\left(1-\rho_{zx}^2\right)\left(1-\lambda C_x^2\right)}{\lambda C_z^2\left(1-\rho_{zx}^2\right)+\left(1-\lambda C_x^2\right)} - \lambda\mu_z^2 C_z^2\left(1-\rho_{zx}^2\right) + \frac{\frac{\mathrm{Var}\left(\hat{\mu}_{DP}\right)^2_{\min}}{\mu_z^2}}{1+\frac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}}{\mu_z^2}} + \frac{\lambda^2\mu_z^2\gamma^2 C_x^2\left\{4C_z^2\left(1-\rho_{zx}^2\right)+\gamma^2 C_x^2\right\}}{4\left\{1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right\}} \\
&= \frac{\frac{4\,\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)^2_{\min}}{\mu_z^2}\left(1+\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}\right)+\left(\lambda C_x^2-1-\frac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}}{\mu_z^2}\right)\left(2\frac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}}{\mu_z}+\lambda\gamma^2\mu_z C_x^2\right)^2}{4\left(1+\frac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}}{\mu_z^2}\right)^2\left(\lambda C_x^2-1-\frac{\mathrm{Var}\left(\hat{\mu}_{\mathrm{DP}}\right)_{\min}}{\mu_z^2}\right)}
\end{aligned}$$

and so $\mathrm{MSE}\left(\hat{\mu}_{\mathrm{GRR}}\right)_{\min} - \mathrm{MinMSE}\left(\hat{\mu}_{\mathrm{NH(exp)}}\right) > 0$ provided that $\left\{\lambda C_x^2 - \lambda C_z^2\left(1-\rho_{zx}^2\right)\right\} > 1$.

Simulation Study

A simulation study is presented to show the performance of the proposed estimator in comparison to other estimators using the auxiliary variable for RRT models. The proposed estimator $\hat{\mu}_{\mathrm{NH(exp)}}$ is compared with the Diana and Perri (2011) estimator $\hat{\mu}_{\mathrm{DP}}$, the Sousa et al. (2010) estimator $\hat{\mu}_{\mathrm{SR}}$, and the Gupta et al. (2012) estimator $\hat{\mu}_{\mathrm{GRR}}$. Three finite populations of size 1000 are generated from a multivariate normal distribution. The three populations each have theoretical mean μ = [5, 5] of [*Y*, *X*] and have different covariance matrices. The populations are generated based on correlation levels between the variables. The correlation levels are classified as low, medium and high. The covariance matrices and the correlations are presented below. The scrambling variable *W* is considered to be a normal random variable with mean equal to zero and standard deviation equal to 0.30. The scrambling variable *T* is considered to be a normal random variable with mean equal to zero and standard deviation equal to 0.20. We use the simulation studies of Gupta et al. to determine the parameters that are easier to compare.

The covariance matrices and the correlation coefficients for each population are given below:

Population I (Low Correlation):

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 9.0 & 5.4 \\ 5.4 & 4.0 \end{bmatrix}, \quad \rho_{yx} = 0.30$$

Population II (Medium Correlation):

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 9.0 & 3.6 \\ 3.6 & 4.0 \end{bmatrix}, \quad \rho_{yx} = 0.60$$

Population III (High Correlation):

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 9.0 & 5.4 \\ 5.4 & 4.0 \end{bmatrix}, \quad \rho_{yx} = 0.90$$

Table 1. Theoretical and empirical MSEs of estimators according to degree of the correlation between the sensitive and non-sensitive variable for model 1 (M_1)

			Estimators			
Population	***n***	**MSE**	$\hat{\mu}_{DP}$	$\hat{\mu}_{SR}$	$\hat{\mu}_{GRR}$	$\hat{\mu}_{NH(exp)}$
I	50	Theoretical	0.1684	0.1894	0.1672	**0.1601**
ρ_{yx} = 0.30		Empirical	0.1728	0.1916	0.1754	**0.1705**
	100	Theoretical	0.0838	0.0942	0.0828	**0.0782**
		Empirical	0.0840	0.0945	0.0840	**0.0808**
	200	Theoretical	0.0415	0.0466	0.0414	**0.0398**
		Empirical	0.0408	0.0465	0.0411	**0.0388**
	300	Theoretical	0.0274	0.0308	0.0273	**0.0268**
		Empirical	0.0272	0.0308	0.0272	**0.0262**
II	50	Theoretical	0.1197	0.1203	0.1191	**0.0972**
ρ_{yx} = 0.60		Empirical	0.1187	0.1191	0.1187	**0.0982**
	100	Theoretical	0.0595	0.0599	0.0594	**0.0494**
		Empirical	0.0608	0.0613	0.0610	**0.0498**
	200	Theoretical	0.0295	0.0296	0.0291	**0.0239**
		Empirical	0.0300	0.0302	0.0297	**0.0244**
	300	Theoretical	0.0194	0.0196	0.0194	**0.0162**
		Empirical	0.0193	0.0193	0.0193	**0.0158**
III	50	Theoretical	0.0358	0.0472	0.0358	**0.0058**
ρ_{yx} = 0.90		Empirical	0.0372	0.0480	0.0374	**0.0060**
	100	Theoretical	0.0178	0.0235	0.0178	**0.0098**
		Empirical	0.0186	0.0239	0.0186	**0.0100**
	200	Theoretical	0.0088	0.0116	0.0088	**0.0033**
		Empirical	0.0091	0.0120	0.0091	**0.0014**
	300	Theoretical	0.0058	0.0077	0.0058	**0.0010**
		Empirical	0.0060	0.0079	0.0060	**0.0010**

Table 2. Theoretical and empirical MSEs of estimators according to degree of the correlation between the sensitive and non-sensitive variable for model 2 (M_2)

			Estimators			
Population	***n***	**MSE**	$\hat{\mu}_{DP}$	$\hat{\mu}_{SR}$	$\hat{\mu}_{GRR}$	$\hat{\mu}_{NH(exp)}$
I	50	Theoretical	3.5609	3.5628	3.1889	**3.1872**
ρ_{yx} = 0.30		Empirical	2.6095	2.5585	2.2517	**2.2515**
	100	Theoretical	1.6867	1.6877	1.5985	**1.5980**
		Empirical	1.5985	1.6009	1.3017	**1.3015**
	200	Theoretical	0.7497	0.7501	0.7317	**0.7316**
		Empirical	1.0440	1.0409	0.8691	**0.8622**
	300	Theoretical	0.4373	0.4376	0.4311	**0.4310**
		Empirical	0.8338	0.8312	0.7260	**0.7259**

Table 2, continued.

			Estimators			
Population	***n***	**MSE**	$\hat{\mu}_{DP}$	$\hat{\mu}_{SR}$	$\hat{\mu}_{GRR}$	$\hat{\mu}_{NH(exp)}$
II	50	Theoretical	3.1069	3.1073	2.8349	**2.8334**
ρ_{yx} = 0.60		Empirical	2.1729	2.1638	1.9602	**1.9517**
	100	Theoretical	1.4717	1.4719	1.4078	**1.4073**
		Empirical	1.3937	1.3960	1.1353	**1.1316**
	200	Theoretical	0.6541	0.6542	0.6412	**0.6410**
		Empirical	0.9189	0.9192	0.7682	**0.7624**
	300	Theoretical	0.3816	0.3816	0.3772	**0.3770**
		Empirical	0.7939	0.7964	0.6994	**0.6898**
III	50	Theoretical	2.8921	2.9277	2.6571	**2.6557**
ρ_{yx} = 0.90		Empirical	2.0042	2.0091	1.8101	**1.7838**
	100	Theoretical	1.3699	1.3868	1.3150	**1.3145**
		Empirical	1.3401	1.3760	1.0960	**1.0958**
	200	Theoretical	0.6089	0.6164	0.5978	**0.5976**
		Empirical	0.9457	0.9552	0.8018	**0.7990**
	300	Theoretical	0.3552	0.3596	0.3514	**0.3512**
		Empirical	0.8639	0.8687	0.7715	**0.7684**

Table 3. Theoretical and empirical MSEs of estimators according to degree of the correlation between the sensitive and non-sensitive variable for model 3 (M_3)

			Estimators			
Population	***n***	**MSE**	$\hat{\mu}_{DP}$	$\hat{\mu}_{SR}$	$\hat{\mu}_{GRR}$	$\hat{\mu}_{NH(exp)}$
I	50	Theoretical	0.1670	0.1883	0.1659	**0.1603**
ρ_{yx} = 0.30		Empirical	0.1722	0.1905	0.1740	**0.1698**
	100	Theoretical	0.0831	0.0937	0.0835	**0.0780**
		Empirical	0.0835	0.0938	0.0847	**0.0798**
	200	Theoretical	0.0411	0.0464	0.0411	**0.0394**
		Empirical	0.0405	0.0462	0.0408	**0.0382**
	300	Theoretical	0.0271	0.0306	0.0271	**0.0257**
		Empirical	0.0270	0.0307	0.0271	**0.0257**
II	50	Theoretical	0.1183	0.1191	0.1178	**0.0964**
ρ_{yx} = 0.60		Empirical	0.1188	0.1180	0.1173	**0.0978**
	100	Theoretical	0.0589	0.0593	0.0588	**0.0484**
		Empirical	0.0606	0.0605	0.0600	**0.0488**
	200	Theoretical	0.0291	0.0293	0.0294	**0.0238**
		Empirical	0.0297	0.0298	0.0300	**0.0242**
	300	Theoretical	0.0192	0.0194	0.0197	**0.0157**
		Empirical	0.0191	0.0191	0.0193	**0.0157**

Table 3, continued.

Population	*n*	MSE	Estimators			
			$\hat{\mu}_{DP}$	$\hat{\mu}_{SR}$	$\hat{\mu}_{GRR}$	$\hat{\mu}_{NH(exp)}$
III	50	Theoretical	0.0357	0.0459	0.0357	**0.0056**
ρ_{yx} = 0.90		Empirical	0.0375	0.0466	0.0370	**0.0061**
	100	Theoretical	0.0178	0.0229	0.0178	**0.0096**
		Empirical	0.0188	0.0233	0.0186	**0.0100**
	200	Theoretical	0.0088	0.0113	0.0088	**0.0031**
		Empirical	0.0090	0.0116	0.0090	**0.0017**
	300	Theoretical	0.0058	0.0075	0.0058	**0.0028**
		Empirical	0.0059	0.0077	0.0060	**0.0028**

Table 4. Theoretical and empirical MSEs of estimators according to degree of the correlation between the sensitive and non-sensitive variable for model 4 (M_4)

Population	*n*	MSE	Estimators			
			$\hat{\mu}_{DP}$	$\hat{\mu}_{SR}$	$\hat{\mu}_{GRR}$	$\hat{\mu}_{NH(exp)}$
I	50	Theoretical	2.0340	2.8564	1.8945	**1.8568**
ρ_{yx} = 0.30		Empirical	2.6028	3.3407	2.2134	**1.9528**
	100	Theoretical	0.9635	1.3530	0.9327	**0.9223**
		Empirical	1.4757	1.6144	1.2717	**1.1878**
	200	Theoretical	0.4282	0.6013	0.4222	**0.4199**
		Empirical	0.8659	0.8751	0.7673	**0.7651**
	300	Theoretical	0.2498	0.3508	0.2478	**0.2469**
		Empirical	0.6889	0.6764	0.6284	**0.6265**
II	50	Theoretical	1.6396	2.4315	1.5521	**1.5146**
ρ_{yx} = 0.60		Empirical	2.2175	2.9569	1.8395	**1.6508**
	100	Theoretical	0.7767	1.1518	0.7577	**0.7479**
		Empirical	1.3376	1.6164	1.1199	**1.0010**
	200	Theoretical	0.3452	0.5119	0.3415	**0.3394**
		Empirical	0.8464	0.9276	0.7488	**0.7016**
	300	Theoretical	0.2014	0.2990	0.2001	**0.1993**
		Empirical	0.7073	0.7303	0.6484	**0.6456**
III	50	Theoretical	1.3325	2.2892	1.2748	**1.2421**
ρ_{yx} = 0.90		Empirical	1.7606	2.1009	1.4869	**1.4323**
	100	Theoretical	0.6312	1.0843	0.6187	**0.6104**
		Empirical	1.1714	1.2880	1.0017	**0.9817**
	200	Theoretical	0.2806	0.4820	0.2781	**0.2763**
		Empirical	0.7664	0.7792	0.6878	**0.6821**
	300	Theoretical	0.1637	0.2811	0.1628	**0.1622**
		Empirical	0.6670	0.6459	0.6185	**0.6139**

The process was repeated 5000 times and for different sample sizes: n = 50, 100, 200, and 300. The value of the design parameter P changes from 0.10 to 0.90 with an increment of 0.1. We observe small differences in efficiency with almost each value of the design parameter when an auxiliary variable is utilized in RRT models. Thus, simulation results are only presented for P = 0.20. That means 20 percent of the respondents gave direct answers; the rest of the respondents use the randomized devices. The performances of the estimators are measured by the simulated MSE:

$$\mathrm{MSE}\left(\hat{\mu}\right)=\frac{1}{5000}\sum_{i=1}^{5000}\left(\hat{\mu}_i-\mu_y\right)^2$$

where $\hat{\mu}_i$ is the estimate of μ_y on the i^{th} sample. Simulation results are summarized in Tables 1-4.

In Tables 1-4, theoretical and empirical MSE values of the estimators, according to degree of the correlation between the sensitive and non-sensitive variables, are given for the four specified models. In all circumstances, regardless of both degree of correlation and sample size, the proposed estimator is always more efficient than the Diana and Perri (2011) estimator $\hat{\mu}_{\mathrm{DP}}$, the Sousa et al. (2010) estimator $\hat{\mu}_{\mathrm{SR}}$, and the Gupta et al. (2012) estimator $\hat{\mu}_{\mathrm{GRR}}$. The MSE values of the estimators are smaller when the sample size increases, and that is an expected result. However, additive models performed better than multiplicative models. When additive models are applied in RRT, more efficient estimates are obtained.

Application

To test the models and show the performance of the proposed estimator in comparison to other estimators, a survey was performed at the Hacettepe University Department of Statistics to estimate the grade point average (GPA) of students who graduated in 2016. One hundred and two students who graduated in 2016 are considered as our population. In this application, the study variable Y is the GPA of students, the auxiliary variable X is study hours per week. Four models for P = 0.20 were applied to the population. Twenty students were requested to report their true GPA, and 82 students used the randomized devices. To apply the randomized devices, random numbers were generated for scrambling variables W and T. For scrambling variable W, 82 random numbers were

generated from the normal distribution with mean equal to zero and standard deviation equal to 0.60. For scrambling variable *T*, 82 random numbers were generated from the normal distribution with mean equal to zero and standard deviation equal to 0.20.

The following are some characteristics of the population:

$$\bar{Y} = 2.51,\ \bar{X} = 7.16,\ S_y^2 = 0.1166,\ S_x^2 = 38.53,\ \rho_{yx} = 0.71$$

Table 5. Theoretical Bias and MSE values of the estimators by using non-sensitive auxiliary variable according to Models

		***n* = 50**		***n* = 100**		***n* = 200**	
Model	**Estimators**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
M_1	$\hat{\mu}_{DP}$	--	0.0155	--	0.0091	--	0.0039
	$\hat{\mu}_{SR}$	0.0563	0.1205	0.0329	0.0705	0.0143	0.0305
	$\hat{\mu}_{GRR}$	0.0061	0.0155	0.0035	0.0091	0.0015	0.0039
	$\hat{\mu}_{NH(exp)}$	0.0041	**0.0151**	0.0024	**0.0089**	0.0011	**0.0038**
M_2	$\hat{\mu}_{DP}$	--	2.3298	--	1.3638	--	0.5909
	$\hat{\mu}_{SR}$	0.1290	2.3519	0.0151	1.3767	0.0054	0.5966
	$\hat{\mu}_{GRR}$	0.5661	1.9313	0.3571	1.2184	0.0337	0.5622
	$\hat{\mu}_{NH(exp)}$	0.5548	**1.9212**	**0.3504**	1.2134	0.0330	**0.5609**
M_3	$\hat{\mu}_{DP}$	--	0.0143	--	0.0084	--	0.0091
	$\hat{\mu}_{SR}$	0.0604	0.1270	0.0354	0.0743	0.0153	0.0322
	$\hat{\mu}_{GRR}$	0.0056	0.0143	0.0033	0.0084	0.0014	0.0036
	$\hat{\mu}_{NH(exp)}$	0.0039	**0.0139**	0.0023	**0.0082**	0.0010	**0.0035**
M_4	$\hat{\mu}_{DP}$	--	2.0248	--	1.1852	--	0.5136
	$\hat{\mu}_{SR}$	0.0474	2.0475	0.0277	1.1985	0.0121	0.5194
	$\hat{\mu}_{GRR}$	0.4974	1.6970	0.3135	1.0696	0.1440	0.4914
	$\hat{\mu}_{NH(exp)}$	0.9625	**0.3203**	0.2991	**0.5629**	0.0276	**0.3902**

Note: Blank cells indicate unbiased estimators.

To compute the Bias and MSE values of the Diana and Perri (2011) estimator $\hat{\mu}_{\mathrm{DP}}$, the Sousa et al. (2010) estimator $\hat{\mu}_{\mathrm{SR}}$, the Gupta et al. (2012) estimator $\hat{\mu}_{\mathrm{GRR}}$, and the proposed estimator $\hat{\mu}_{\mathrm{NH(exp)}}$ for the four models based on different sample sizes: n = 20, 30, and 50, arbitrarily take $\alpha = 1$ and $\beta = -1$, that is

$$\gamma = \frac{\mu_x}{2\left(\mu_x - 1\right)}$$

for simplicity. The results are summarized in Table 5.

In the application study, the most efficient estimator was the proposed exponential-type estimator. It was always more efficient than the existing estimators in all RRT models for different sample sizes. From Table 5, it can be concluded that the additive models were more efficient than the multiplicative models and that the proposed estimator gave better results.

Conclusion

An exponential-type estimator was proposed, based on a non-sensitive auxiliary variable, for the population mean of a sensitive variable for Generalized Quantitative RRT models. The MSE equation is derived for all Quantitative RRT models. The proposed estimator was more efficient than other existing estimators in all circumstances, regardless of which model was applied. It was shown that the efficiency of the proposed estimator can be quite substantial if the correlation between the study and the auxiliary variables is high. Additionally, the additive models were more efficient than the multiplicative models. These results were supported by simulation and application studies. In a future work, an estimator will be developed for the population mean of the sensitive study variable by combining additive and multiplicative techniques based on Quantitative RTT using multi-sensitive auxiliary variables.

References

Bahl, S., & Tuteja, R. K. (1991). Ratio and product type exponential estimators. *Journal of Information and Optimization Sciences, 12*(1), 159-164. doi: 10.1080/02522667.1991.10699058

Bar-Lev, S. K., Bobovitch, E., & Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika, 60*(3), 255-260. doi: 10.1007/s001840300308

Cingi, H., & Kadilar, C. (2009). *Advances in sampling theory – Ratio method of estimation*. Bentham Science Publishers. doi: 10.2174/97816080501231090101

Diana, G. & Perri, P. F. (2011). A class of estimators for quantitative sensitive data. *Statistical Papers, 52*(3), 633-650. doi: 10.1007/s00362-009-0273-1

Eichhorn, B. H., & Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, *7*(4), 307-316. doi: 10.1016/0378-3758(83)90002-2

Grover, L. K., & Kaur, P. (2014). A generalized class of ratio type exponential estimators of population mean under linear transformation of auxiliary variable. *Communications in Statistics – Simulation and Computation, 43*(7), 1552-1574. doi: 10.1080/03610918.2012.736579

Gupta, S., Shabbir, J., Sousa, R., & Real, P. C. (2012). Estimation of the mean of a sensitive variable in the presence of auxiliary information. *Communications in Statistics – Theory and Methods, 41*(13-14), 1-12. doi: 10.1080/03610926.2011.641654

Özgül, N. (2013). *Proportion and mean estimators in randomized response models* (Unpublished doctoral thesis). Hacettepe University, Ankara, Turkey.

Özgül, N. & Cingi, H. (2014). A new class of exponential regression cum ratio estimator in two phase sampling. *Hacettepe Journal of Mathematics and Statistics, 43*(1), 131-140. Available from http://dergipark.ulakbim.gov.tr/hujms/article/view/5000017145

Shabbir, J., & Gupta, S. (2011). On estimating finite population mean in simple and stratified random sampling. *Communications in Statistics – Theory and Techniques, 40*(2), 199-212. doi: 10.1080/03610920903411259

Sousa, R., Shabbir, J., Real, P. C., & Gupta, S. (2010). Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, *4*(3), 495-507. 10.1080/15598608.2010.10411999

Thornton, B., & Gupta, S. N. (2004). Comparative validation of a partial (versus full) randomized response technique: Attempting to control for social

desirability response bias to sensitive questions. *Individual Differences Research, 2*(3), 214-224.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*(309), 63-69. doi: 10.2307/2283137

Warner, S. L. (1971). The linear randomized response model. *Journal of the American Statistical Association, 66*(336) 884-888. doi: 10.2307/2284247

Appendix A: Special Models for Generalized RTT

***First Model* (M_1):** $S = Y + W$, $R = X$, Mean given by

$$\mu_z = \mu_y + (1-P)\mu_w$$
$$\mu_u = \mu_x$$
$$c = (1-P)\mu_w,\ h = 1$$

Variance and correlation equations to be used in (17) are given by

$$S_z^2 = S_y^2 + (1-P)\mu_w^2\left(C_w^2 + P\right)$$
$$S_u^2 = S_x^2$$
$$\rho_{zu} = \frac{S_{yx}}{S_x\sqrt{S_y^2 + (1-P)\mu_w^2\left(C_w^2 + P\right)}}$$

***Second Model* (M_2):** $S = YW$, $R = X$, Mean given by

$$\mu_z = \left\{P + (1-P)\mu_w\right\}\mu_y$$
$$\mu_u = \mu_x$$
$$c = 0,\ h = \left\{P + (1-P)\mu_w\right\}$$

Variance and correlation equations to be used in (17) are given by

$$S_z^2 = \mu_y^2\left(1+C_y^2\right)\left\{P+\left(1-P\right)\mu_w^2\left(1+C_w^2\right)\right\}-\mu_z^2$$
$$S_u^2 = S_x^2$$
$$\rho_{zu} = \frac{S_{yx}\left\{P+\left(1-P\right)\mu_w\right\}}{S_x\sqrt{\mu_y^2\left(1+C_y^2\right)\left\{P+\left(1-P\right)\mu_w^2\left(1+C_w^2\right)\right\}-\mu_z^2}}$$

***Third Model* (M3):** $S = Y + W$, $R = X + T$, Mean given by

$$\mu_z = \mu_y + \left(1-P\right)\mu_w$$
$$\mu_u = \mu_x + \left(1-P\right)\mu_t$$
$$c = \left(1-P\right)\mu_w,\ h = 1$$

Variance and correlation equations to be used in (17) are given by

$$S_z^2 = S_y^2 + \left(1-P\right)\mu_w^2\left(C_w^2+P\right)$$
$$S_u^2 = S_x^2 + \left(1-P\right)\mu_t^2\left(C_t^2+P\right)$$
$$\rho_{zu} = \frac{S_{yx}+P\left(1-P\right)\mu_w\mu_t}{S_x\sqrt{\left\{S_y^2+\left(1-P\right)\mu_w^2\left(C_w^2+P\right)\right\}\left\{S_x^2+\left(1-P\right)\mu_t^2\left(C_t^2+P\right)\right\}}}$$

Fourth Model* (M4)*: *S* = *YW*, *R* = *XT*, Mean given by

$$\mu_z = \left\{P + (1-P)\,\mu_w\right\}\mu_y$$
$$\mu_u = \left\{P + (1-P)\,\mu_t\right\}\mu_x$$
$$c = 0,\ h = P + (1-P)\,\mu_w$$

Variance and correlation equations to be used in (17) are given by

$$S_z^2 = \mu_y^2\left(1-C_y^2\right)\left[P + (1-P)\,\mu_w^2\left(1+C_w^2\right)\right] - \mu_z^2$$
$$S_u^2 = \mu_x^2\left(1+C_x^2\right)\left[P + (1-P)\,\mu_t^2\left(1+C_t^2\right)\right] - \mu_u^2$$
$$\rho_{zu} = \frac{\left\lfloor \sigma_{yx}\left\{P + (1-P)\,\mu_t\mu_w\right\} + P(1-P)\,\mu_y\mu_x\left(1-\mu_w\right)\left(1-\mu_t\right)\right\rfloor}{\sqrt{\left[\mu_y^2\left(1-C_y^2\right)\left[P + (1-P)\,\mu_w^2\left(1+C_w^2\right)\right] - \mu_z^2\right]\left[\mu_x^2\left(1+C_x^2\right)\left[P + (1-P)\,\mu_t^2\left(1+C_t^2\right)\right] - \mu_u^2\right]}}$$

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 388-405.
doi: 10.22237/jmasm/1493598120

ISSN 1538 − 9472

A Comparison of Depth Functions in Maximal Depth Classification Rules

Olusola Samuel Makinde
Federal University of Technology
Akure, Nigeria

Adeyinka Damilare Adewumi
Federal University of Technology
Akure, Nigeria

Data depth has been described as alternative to some parametric approaches in analyzing many multivariate data. Many depth functions have emerged over two decades and studied in literature. In this study, a nonparametric approach to classification based on notions of different data depth functions is considered and some properties of these methods are studied. The performance of different depth functions in maximal depth classifiers is investigated using simulation and real data with application to agricultural industry.

Keywords: classification rules, data depth, error rates, non-parametric approach, symmetry

Introduction

Classification is a practical subject in statistics. It aims at assigning an unclassified observation to one of several groups or populations on the basis of some measurement. Anderson (1984) described classification problem as a problem of statistical decision-making. However, classical multivariate analysis has relied heavily on the assumption of normality in data presentation and analysis. Among the classification methods that rely heavily on distribution assumption are Bayes rule (Welch, 1938), linear discriminant analysis and quadratic linear discriminant analysis (Anderson, 1984), and independence rule (Dudoit, Fridlyand & Speed, 2002). Research has shown that most of the data acquired nowadays do not satisfy normality assumption. Similarly, some parametric approaches are prone to the effect of outlying observations. This gives nonparametric approach to classification an edge over parametric methods.

Olusola Samuel Makinde is a Lecturer in the Department of Statistics. Email at osmakinde@futa.edu.ng. Adeyinka Damilare Adewumi is a graduate of the Department of Statistics. Email at adewumiadeyinkad@yahoo.com.

Other methods in literature include support vector machine (Vapnik, 1998; Cortes & Vapnik, 1995), nearest neighbour rule (Cover & Hart, 1967), classification rules based on distance functions (Chan & Hall, 2009; Hall, Titterington & Xue, 2009), classifiers based on distribution functions of rank outlyingness (Makinde & Chakraborty, 2015).

Data depth is a way to measure the depth or outlyingness of a given point with respect to a multivariate data cloud or its underlying distribution (Liu, Singh & Parelius, 1999). It gives rise to a natural centr-outward ordering of the sample points in $\mathbb{R}^d$. This ordering gives rise to new and easy ways to quantify many complex multivariate features of the underlying distribution, including location, quantiles, scale, skewness and kurtosis. Liu (1990) introduced a notion of simplicial depth and corresponding estimators of location, and formulated a quality index with simplicial depth, Mahalanobis depth and majority depth. Koshevoy & Mosler (1997) introduced a notion of zonoid depth while Fraiman, Meloche & García-Escudero (1999) introduced a likelihood type depth function. Rousseeuw & Hubert (1999) introduced a notion of regression depth. Liu, Singh & Parelius (1999) considered some examples of depth functions and developed methodology for their practical applications.

Classification rule based on data depth is considered in the current study. Data depth is formally defined based on Zuo & Serfling (2000a) and examples of depth functions are presented. In reality, an important question that arises in almost all fields where supervised learning is employed is that which of the depth functions should be employed. Classification rules based on the depth functions are defined and properties of the classification rules are presented. Evaluation of the classification rule, accounting for performance of various depth functions are presented based on numerical examples.

Notions of Statistical Depth Functions

Definition 1 (Zuo & Serfling, 2000a). Let the mapping $D(.;.) : \mathbb{R}^d \times \boldsymbol{\mathcal{F}} \to \mathbb{R}$ be bounded and non-negative, and satisfy:

i. $D(\mathbf{Ax} + \mathbf{b}, F_{\mathbf{AX}+\mathbf{b}}) = D(\mathbf{x}, F\mathrm{x})$ holds for any random vector $\mathbf{X} \in \mathbb{R}^d$ and any $d \times d$ nonsingular matrix $\mathbf{A}$, and any d dimensional vector $\mathbf{b}$.

ii. $D(\boldsymbol{\theta},F) = \sup_{\mathbf{x}\in\mathbb{R}^d} D(\mathbf{x},F)$ holds for any $F \in \boldsymbol{\mathcal{F}}$ having centre $\boldsymbol{\theta}$.

iii. For any $F \in \boldsymbol{\mathcal{F}}$ having deepest point $\boldsymbol{\theta}$, $D(\mathbf{x},F) \leq D(\boldsymbol{\theta} + \alpha(\mathbf{x} - \boldsymbol{\theta}),F)$ holds for $\alpha \in [0,1]$; and

iv. $D(\mathbf{x},F) \to 0$ as $\|\mathbf{x}\| \to \infty$, for each $F \in \mathcal{F}$.

Then $D(.,F)$ is called a statistical depth function.

From Definition 1, the first property describes invariance of depth function under general affine transformation of the data. That is, the depth of any observation in $\mathbb{R}^d$ should not depend on the scale of the underlying measurement or underlying coordinate system. The second property implies that depth value attains its maximum value at the point of symmetry for symmetric distributions. The third property implies that the depth value decreases monotonically as vector **x** moves away from its most central point while the fourth property implies that the depth value of **x** vanishes (tend to zero) as Euclidean norm of **x** approaches infinity.

The depth functions in literature include

1. ***Mahalanobis Depth (MhD)***. Mahalanobis (1936); Liu & Singh (1993) defined the depth of an observation **x** with respect to the distribution F as

$$MhD(\mathbf{x},F) = [1 + O(\mathbf{x},\boldsymbol{\mu}_F,\Sigma_F)]^{-1}$$

where $O(\mathbf{x},\boldsymbol{\mu}_F,\Sigma_F) = (\mathbf{x} - \boldsymbol{\mu}_F)' \, \Sigma_F^{-1}(\mathbf{x} - \boldsymbol{\mu}_F)$, $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ are the mean vector and dispersion matrix of F respectively. The sample version of *MhD* is obtained by replacing $\boldsymbol{\mu}_F$ and Σ_F with their estimates.

2. ***Zonoid Depth (ZD).*** Dyckerhoff et al. (1996) defined a zonoid depth as

$$ZD(\mathbf{x},F) = \sup\{\alpha : \mathbf{x} \in D_\alpha(\mathbf{X}_1,\ldots,\mathbf{X}_n)\}$$

where $D_\alpha(\mathbf{X}_1,\ldots,\mathbf{X}_n) = \Sigma_{i=1}^n \lambda_i \mathbf{X}_i$, $\Sigma_{i=1}^n \lambda_i = 1$, $\lambda_i \geq 0$, and $\alpha\lambda_i \leq \frac{1}{n}$ for all i.

3. ***Half-Space Depth (HD).*** Tukey (1975) defined half-space depth of a point $\mathbf{x} \in \mathbb{R}^d$ with respect to F as the minimum probability mass carried by any closed half-space containing **x**, Mathematically,

$$HD(\mathbf{x},P) = \inf_H [P(H)]$$

where H is a closed halfspace in R^d and $\mathbf{x} \sim H$.

4. ***Oja Depth (OD).*** Oja (1983) defined the depth of $\mathbf{x} \in \mathbb{R}^d$ with respect to F as

$$OD(F;\mathbf{x}) = [1 + O(\mathbf{x},F)]^{-1}$$

where $O(\mathbf{x},F) = E_F(\text{Volume}(S[\mathbf{x}, \mathbf{X}_1,\ldots,\mathbf{X}_d]))$, $S[\mathbf{x}, \mathbf{X}_1,\ldots,\mathbf{X}_d]$ is a closed simplex with vertices $\mathbf{x}$ and d random observations $\mathbf{X}_1,\ldots,\mathbf{X}_d$ from F.

5. ***Simplicial Depth (SD).*** Liu (1990) defined simplicial depth of $\mathbf{x} \in \mathbb{R}^d$ with respect to F as

$$SD(F;\mathbf{x}) = \mathrm{P}(\mathbf{x} \in S[\mathbf{x}, \mathbf{X}_1,\ldots,\mathbf{X}_{d+1}])$$

where $S[\mathbf{x}, \mathbf{X}_1,\ldots,\mathbf{X}_{d+1}]$ is a closed simplex formed by $(d + 1)$ random observation from F. The sample version of $SD(F;\mathbf{x})$ is obtained by replacing F in $SD(F;\mathbf{x})$ by F_n.

6. ***Projection Depth (PD).*** Donoho & Gasko (1982) defined the depth of $\mathbf{x}$ with respect to F as the worst case outlyingness of $\mathbf{x}$ with respect to one dimensional median in any one-dimensional projection.

$$PD(F;\mathbf{x}) = (1 + O(\mathbf{x},F))^{-1}$$

where $O(\mathbf{x}, F) = \sup_{\|u\|=1} \frac{\mathbf{u}'\mathbf{x} - Med(F_u)}{MAD(F_u)}$, F_u is the distribution $\mathbf{u}'\mathbf{X}$, $Med(F_u)$ is the median of F_u, $MAD(F_u)$ is the median absolute deviation of F_u and $X \sim F$. The sample version of $PD(F;\mathbf{x})$ is obtained by replacing the median and MAD with their sample estimates.

7. ***Likelihood Depth (LD).*** Fraiman, Meloche & García-Escudero (1999) defined the depth of $\mathbf{x}$ with respect to F simply as its probability density, that is, $LD(F;\mathbf{x}) = f(\mathbf{x})$, and the empirical version

can be any consistent density estimate at **x**, for example, the kernel density estimate.

8. ***Spatial Depth (SPD).*** Serfling (2002) defined spatial depth of any observation **x** with respect to F as

$$SBD(\mathbf{x}, F) = 1 - \left\| E_F \left(\frac{\mathbf{x} - \mathbf{X}}{\|\mathbf{x} - \mathbf{X}\|} \right) \right\|$$

where $X \sim F$.

9. ***Simplicial volume depth (SVD)***. Zuo & Serfling (2000a, b) expressed SVD of an observation **x** with respect to F as

$$SVD(\mathbf{x}, \delta, F) = \left\{ 1 + E_F \left(\frac{\nabla(\mathbf{x}, \mathbf{X}_1, ..., \mathbf{X}_d)}{|\Sigma_F|^{\frac{1}{2}}} \right)^{\delta} \right\}^{-1}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_d$ are independent and identically distributed observations from F, $\nabla(\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_d)$ is the volume of the d-dimensional simplex formed by **x** and Σ_F is the scatter matrix of the distribution F.

10. ***Majority Depth (MJD)***. Liu & Singh (1993) defined the depth of **x** with respect to F as the probability that **x** belongs to the major side (i.e. the half-space with the larger probability measure) of a random hyperplane passing through the data points in $\mathbb{R}^d$.

Other depth functions include regression depth (Rousseeuw & Hubert, 1999). Gao (2003) defined another depth function based on square of spatial outlyingness function. Few of these depth functions satisfy all the four properties in Definition 1 while others satisfy some of the properties. See Zuo & Serfling (2000a; 2000b) for detail.

Classification Rule

The goal of any classification rule is to find a rule or tool that enables us to assign an observation $\mathbf{x} \in \mathbb{R}^d$ to one of the several competing groups (or classes). One can define a classification rule based on depth functions. It is easy to observe that data depth gives an idea on how outlying an observation $\mathbf{x}$ is with respect to the distribution F. If $\mathbf{x}$ is a central observation, its depth value will be large. On the other hand, if $\mathbf{x}$ is an extreme observation, its depth value will be small. Thus a small depth value may suggest a deviation of $\mathbf{x}$ from F.

Ghosh & Chaudhuri (2005) proposed a classification rule based on simple idea of assigning a new observation to any of the J competing classes, for which it attains maximal depth value. It is expressed as:

$$D\left(F_k, \mathbf{x}\right) = \arg\max_{1 \le j \le J} D\left(F_j, \mathbf{x}\right) \tag{1}$$

where F_k is the distribution of k^{th} class and $1 \le j \le J$.

Let us consider two classes for simplicity. Suppose π_j has multivariate distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, $j = 1,2$. For $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the classification rule in (1) can be expressed as

$$\text{Assign } \mathbf{x} \text{ to } F \text{ if } D\left(F, \mathbf{x}\right) > D\left(G, \mathbf{x}\right), \text{ and to } G \text{ if otherwise.} \tag{2}$$

It is straightforward to show that a depth function can be expressed in terms of probability density function of the competing distribution. This result is presented by a Lemma below:

Lemma 1. Let F_j be spherically symmetric distributions with density functions of the form

$$f_j\left(\mathbf{x}\right) = \left|\boldsymbol{\Sigma}_j\right|^{-\frac{1}{2}} h\left(\mathbf{x} - \boldsymbol{\theta}_j\right)' \left(\mathbf{x} - \boldsymbol{\theta}_j\right)$$

for some strictly decreasing, continuous, non-negative scalar function h. Then for any of the depth functions OD and SPD,

$$f_j\left(\mathbf{x}\right) = \omega\left(D\left(F_j, \mathbf{x}\right)\right)$$

for some increasing function ω.

Suppose a random vector $\mathbf{X}$ in $\mathbb{R}^d$ is elliptically distributed such that its density is of the form $f(\mathbf{x})=|\mathbf{\Sigma}|^{-½} h((\mathbf{x}-\boldsymbol{\theta})' \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\theta}))$, then $D(F,\mathbf{x})$ can be expressed as a function of $(\mathbf{x}-\boldsymbol{\theta})' \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\theta})$. This result is presented formally by a Lemma below:

Lemma 2. Let F_j be elliptically symmetric distributions with density functions of the form

$$f_j(\mathbf{x}) = |\mathbf{\Sigma}_j|^{-\frac{1}{2}} h\left((\mathbf{x}-\boldsymbol{\theta}_j)' \mathbf{\Sigma}_j^{-1} (\mathbf{x}-\boldsymbol{\theta}_j)\right)$$

for some strictly decreasing, continuous, non-negative scalar function h. Then for any of the depth functions detailed earlier, except OD and SPD,

$$f_j(\mathbf{x}) = \omega\left(D(F_j,\mathbf{x})\right),$$

where $\mathbf{\Sigma}_j$ is not a constant multiple of identity matrix for some increasing function ω.

The optimal rule, Bayes rule, assigns an observation to the class or distribution with highest posterior probability. That is, assign $\mathbf{x}$ to j^{th} class class if $p_j f_j(\mathbf{x})$ is the highest, where p_j is the prior probability of the j^{th} class. Based on the results of Lemmas 1 and 2, it is straightforward to show that maximum depth classifiers are Bayes rules under necessary conditions.

Theorem 1. Suppose the conditions of Lemmas 1 and 2 hold on all the depth functions defined earlier. Then the classifier defined in (1) is Bayes rule if competing distributions have equal covariance matrices and prior probabilities.

In practice, a depth function may not be completely known and so need to be estimated based on sample and then define the empirical version of the classification rule based on the empirical depth function. The empirical depth function based on sample is denoted by $D(F_n,\mathbf{x})$. To show the consistency of empirical depth functions, it is desirable to establish the almost sure convergence of empirical depth functions to its population counterpart.

Theorem 2. Suppose $D(F_n,\mathbf{x})$ is an empirical depth function based on $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Let $D(F,\mathbf{x})$ be a population depth function of any random vector $\mathbf{x}$. Then for any $\mathbf{x}$ in the support of F,

$$\sup_{\mathbf{x}} \left| D(F_n,\mathbf{x}) - D(F,\mathbf{x}) \right| \to 0, \; n \to \infty.$$

The almost sure convergence of half-space depth has been established in Donoho & Gasko (1992), simplicial depth in Liu (1990). Liu & Singh (1993) has shown almost sure convergence of Mahalanobis depth and majority depth while Zuo & Serfling (2000b) proved convergence of projection depth. The almost sure convergence of spatial depth follows from Koltchinskii's (1997) work on the convergence of the empirical spatial rank function to its population version. Convergence of the empirical classification rule to population version follows from Theorem 2.

Evaluation of Classification methods

One way of evaluating the performance of a classifier is to compute its associated misclassification probability. In a two class classification problem, one can define a misclassification probability as

$$\Delta = p_1 P\left(D(F,\mathbf{x}) < D(G,\mathbf{x}) \,|\, \mathbf{x} \in F\right) + p_2 P\left(D(F,\mathbf{x}) > D(G,\mathbf{x}) \,|\, \mathbf{x} \in G\right)$$

The empirical version of the probability of misclassification or error rate, denoted by $\hat{\Delta}$, can be defined as

$$\hat{\Delta} = \frac{p_1}{n} \sum_{i=1}^{n} I\left\{ D\left(\hat{F},\mathbf{x}_i\right) < D\left(\hat{G},\mathbf{x}_i\right) \,|\, \mathbf{x} \in F \right\} + \frac{p_2}{m} \sum_{i=1}^{m} I\left\{ D\left(\hat{F},\mathbf{x}_i\right) > D\left(\hat{G},\mathbf{x}_i\right) \,|\, \mathbf{x} \in F \right\}$$

Under the conditions of Theorems 1 and 2, it is straightforward to show that $\hat{\Delta}$ is a Bayes risk.

Simulation Study

As illustration of the performance of maximum depth classification methods, consider the following example. Let populations π_1 and π_2 be bivariate spherically symmetric with centres of symmetry $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and covariance matrices, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Assume that the prior probabilities of π_1 and π_2 are equal. Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is a random sample from π_1 and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$, a random sample from π_2. New random vectors $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_m$ from π_1 and $\mathbf{Z}_{m+1}, \mathbf{Z}_{m+2}, \ldots, \mathbf{Z}_{2m}$ from π_2 are generated and sample sizes n and m are taken to be 100. $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are chosen to be $(0 \;\; 0)^T$ and $(\delta \;\; 0)^T$ respectively for $\delta \in [-2,2]$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2$. The simulation size is taken to be 1000. Different depth functions are considered for some competing distributions. The distributions are bivariate normal distributions and bivariate Laplace distributions. For computation of likelihood depth, Gaussian kernel is used with turning parameter (=0.3). *R* Package `fda.usc` is used for computing projection and likelihood depth. *R* Package `depth` is used for computing Oja depth, simplicial depth and half-space depth while *R* Package `ddalpha` is used for computing simplicial volume depth, Mahalanobis depth and Zonoid depth.

Estimates of misclassification probabilities are less in bivariate normally distributed samples than bivariate Laplace samples, as shown in Figure 1. It is observed from Figures 2 and 3 that maximal depth classification rule based on half-space depth outperforms others when the distinction between competing distributions is not wide. That is, when $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \to \mathbf{0}$. The distinction between competing distributions becomes clear as $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ moves away from $\mathbf{0}$ and the performance of various depth functions becomes equivalent. It is noted that exact computation of half-space depth and simplicial depth functions is feasible only in $\mathbb{R}^3$ and $\mathbb{R}^2$ respectively. Cuesta-Albertos & Nieto-Reyes (2010) suggested a modified version of half-space depth for functional data, as extension of multivariate set-up. The performance of empirical likelihood depth based on kernel estimator of probability density function depends on the choice of kernel function and turning parameter. It is observed that spatial depth and Oja depth are not invariant under general affine transformation. Makinde (2017) considered various affine invariant versions of spatial rank, a related notion to spatial depth. Robustness of spatial rank (a straightforward extension of spatial depth) against deviation from notion of elliptical symmetry is demonstrated in Makinde and Chakraborty (2015).

Maximum depth classification rule is compared with some classification methods, which include linear discriminant analysis (LDA), *k*-nearest neighbor

rule (kNN) and support vector machine (SVM); using the above setting for δ = 1,2. Table 1 below presents performance of classifiers. It is observed from the table that maximum depth classifiers based on half-space depth has the best performance among the depth based procedures, linear discriminant analysis, *k*-nearest neighbor rule and support vector machine. It has the least mean error rates when the competing distributions are normal and Laplace. Next to half-space depth among the depth functions for maximum depth classification rule is zonoid depth.

However, zonoid depth is not robust against outlying observations in the data cloud. LDA performs well compared with kNN and SVM. It is noted that linear discriminant analysis is Bayes (optimal) rule when competing distributions are multivariate normal. Hence maximum depth classifiers based on half-space depth is a better alternative to the known parametric classification methods, e.g. LDA.

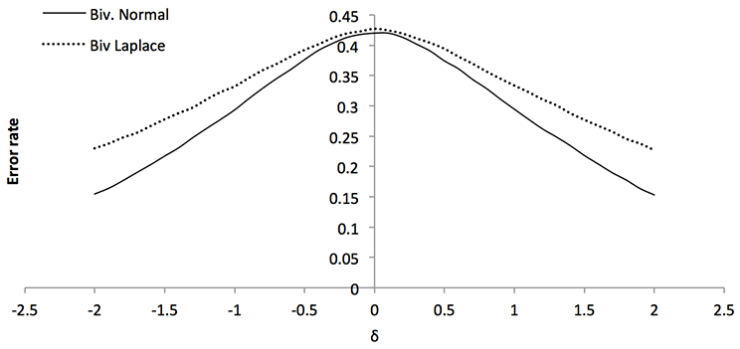


Figure 1. Comparison of error rates associated with half space depth for normally distributed samples and Laplace distributed samples.

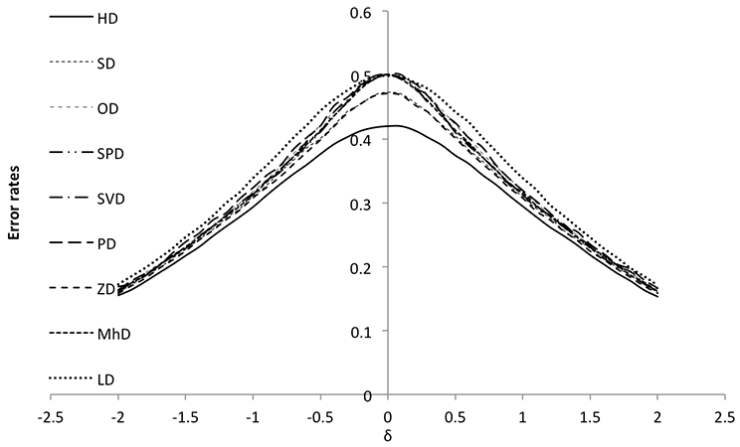


Figure 2. Comparison of depth functions in classification based on error rates for normally distributed samples

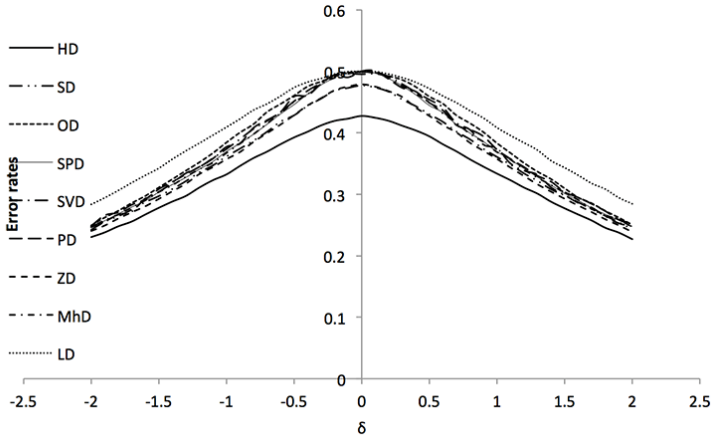


Figure 3. Comparison of depth functions in classification based on error rates Laplace distributed samples.

Table 1. Comparison of mean error rates of classifiers when competing distributions differ in location.

		Maximal Depth Classifiers											
Distribution	**δ**	**HD**	**SD**	**OD**	**SPD**	**SVD**	**PD**	**ZD**	**MhD**	**LD**	**LDA**	**kNN**	**SVM**
Biv normal	1	0.295	0.309	0.318	0.313	0.313	0.319	0.307	0.316	0.334	0.315	0.356	0.316
	2	0.153	0.160	0.165	0.162	0.161	0.167	0.159	0.161	0.167	0.161	0.181	0.167
Biv Laplace	1	0.334	0.361	0.383	0.369	0.375	0.373	0.357	0.369	0.413	0.377	0.410	0.381
	2	0.227	0.243	0.250	0.248	0.248	0.253	0.239	0.246	0.287	0.246	0.273	0.257

Table 2. Comparison of mean error rates of classifiers when competing distributions differ in location and scale.

	Maximal Depth Classifiers											
	HD	**SD**	**OD**	**SPD**	**SVD**	**PD**	**ZD**	**MhD**	**LD**	**QDA**	**kNN**	**SVM**
Biv normal	0.382	0.386	0.500	0.387	0.389	0.386	0.389	0.389	0.166	0.142	0.209	0.148
Biv Laplace	0.410	0.417	0.500	0.418	0.421	0.419	0.421	0.418	0.255	0.214	0.282	0.214

Table 3. Comparison of computation time of classifiers for bivariate Laplace distributions.

	Maximal Depth Classifiers											
	HD	**SD**	**OD**	**SPD**	**SVD**	**PD**	**ZD**	**MhD**	**LD**	**QDA**	**kNN**	**SVM**
Time (seconds)	0.12	0.12	0.14	0.52	15.84	5.97	0.34	0.32	1.39	0.08	0.05	0.31

Only populations which are separated by location are considered so far. Table 2 presents a comparison of proportions of misclassification of depth based procedures, quadratic discriminant analysis (kNN) and SVM when competing populations have different location vectors and covariance matrices.

Suppose the mean vectors and covariance matrices of π_1 and π_2 are $(\boldsymbol{\mu}_1 = (0\ \ 0)^T, \boldsymbol{\Sigma}_1 = \mathbf{I}_2)$ and $(\boldsymbol{\mu}_2 = (2\ \ 0)^T, \boldsymbol{\Sigma}_2 = 9\mathbf{I}_2)$, respectively. It is well known that QDA is an optimal rule when competing populations are normally distributed and differ in location and scale. Hence it has a least mean error rate (= 0.142) for normal distributions. Maximum depth classifier based on likelihood depth has the least mean error rate (= 0.166) among the depth classifiers, which is competitive with QDA and SVM (with mean error rate = 0.148). Maximum depth classifier based on Oja depth has the worst performance in this case. For bivariate Laplace distributions, Maximum depth classifier based on likelihood depth has the least mean error rate (= 0.255) among the depth classifiers, which is competitive with QDA (with mean error rate = 0.214), SVM (with mean error rate = 0.214) and kNN (with mean error rate = 0.282). Mean error rates of other depth classifiers are a bit high.

Presented in Table 3 is a comparison of computation time in seconds of each classifier when competing distributions are bivariate Laplace for one repetition. It is shown in Table 3 that QDA and kNN have the least computation time. However, computation time of maximum depth classifiers based on half-space depth, simplicial depth, Oja depth, zonoid depth and spatial depth are competitive with those of parametric classifiers.

Analysis of Real Data

A real dataset is also analysed to illustrate the performances of depth functions in maximal depth classification methods. Maximal depth classifiers are applied on mineral ions variability data. The data was extracted from a project experiment on crop science and production at the Institute for Agricultural Research and Training (IAR&T) project titled "inter- and intra-maturity group differences in physiological quality of maize seeds" (Olasoji, 2014). The data contains measurements of mean amount of mineral ions (Na, Ca, K and P) leaked after 24 hours from soaked maize seeds at different maturity groups (early, intermediate and lately). Each observation consists of four attributes, which are mean mineral ions (Na, Ca, K and P). Each group consists of 36 observations. A random sample of size 30 and a test sample of size 6 are chosen. The experiment is repeated 100 times; quantile, mean and standard deviation of the proportions of misclassification associated with each of the classifiers are computed.

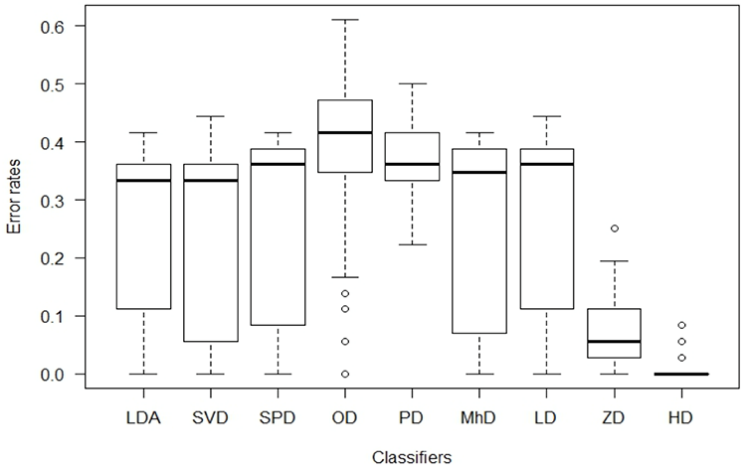


Figure 4. Box plot of proportions of misclassification associated with some classifiers for real data example

Table 4. Quantiles, means and standard deviations of proportions of misclassification of some classifiers for real data example.

	Maximal Depth Classifiers								
	HD	**OD**	**SPD**	**SVD**	**PD**	**ZD**	**MhD**	**LD**	**LDA**
Minimum	0.0000	0.0000	0.0000	0.0000	0.2222	0.0000	0.0000	0.0000	0.0000
25% Quantile	0.0000	0.3541	0.0833	0.0556	0.3333	0.0278	0.0764	0.1111	0.1111
Mean	0.0069	0.3861	0.2697	0.2364	0.3717	0.0725	0.2683	0.2767	0.2619
Median	0.0000	0.4167	0.3611	0.3333	0.3611	0.0556	0.3472	0.3611	0.3333
75% Quantile	0.0000	0.4722	0.3889	0.3611	0.4167	0.1111	0.3889	0.3889	0.3611
Maximum	0.0556	0.6111	0.4444	0.4444	0.5000	0.2500	0.4167	0.4167	0.4167
Standard deviation	0.0150	0.1478	0.1575	0.1618	0.0623	0.0600	0.1533	0.1462	0.1472

Presented in Figure 4 is a comparison of maximum depth classifiers with linear discriminant analysis based on the proportions of misclassification using box plot. The figure shows that the maximum depth classifiers based on half-space depth and zonoid depth has the least proportions of misclassification while the maximum depth classifiers based on Oja depth and projection depth has highest proportions of misclassification.

Presented in Table 4 is the quantile, mean and standard deviation of the proportions of misclassification associated with each of the competing classifiers. Maximum depth classifier based on half-space depth has the least mean proportion of misclassification as shown in the table. Use of spatial depth, simplicial volume depth, Mahalanobis depth and likelihood depth in maximum depth classifiers perform equivalently to LDA, while maximum depth classifiers based on half-space depth and zonoid depth outperform LDA. Simplicial depth values could not be computed as $d = 4 > 2$. For computation of half-space depth, an approximate algorithm implemented in *R* Package `depth` is used.

Conclusion

The maximum depth classifiers based on the training samples when any of the half-space depth, projection depth, simplicial depth, spatial depth, Oja depth, and majority depth is used, do not depend on any distributional assumptions or do not require any estimation of model parameters. That gives maximum depth classifiers an importance over parametric methods. Maximum depth classifiers are easily lent to multiclass cases. We have noted in our real data examples that the maximum depth classifiers are quite competitive with similar classifiers, especially when any of half-space or zonoid depth is used.

Acknowledgment

The authors will like to thank Dr. J. Olasoji of the Institute for Agricultural Research and Training (IAR&T), Ibadan for making mineral ions variability data available.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* NY: John Wiley & Sons, Inc.

Chan, Y. & Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika, 96*(2), 469 – 478. doi: 10.1093/biomet/asp007

Cortes, C. & Vapnik, V. N. (1995). Support-vector networks. *Machine Learning, 20*(3), 273 – 297. doi: 10.1007/bf00994018

Cover, T. M. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21 – 27. doi: 10.1109/tit.1967.1053964

Cuesta-Albertos, J. A. & Nieto-Reyes, A. (2010). Functional classification and the random Tukey depth. Practical issues. *Advances in Intelligent and Soft Computing, 77*, 123 – 130. doi: 10.1007/978-3-642-14746-3_16

Donoho, D. L. & Gasko, M. (1992). Breakdown properties of multivariate location parameters and dispersion matrices. *The Annals of Statistics, 20*(4), 1803 – 1827. doi: 10.1214/aos/1176348890

Dudoit S., Fridlyand, J. & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*(457), 77 – 87. doi: 10.1198/016214502753479248

Dutta, S. & Ghosh, A. K. (2012). *On classification based on L_p depth with an adaptive choice of p.* Technical Report No. R5/2011, Statistics and Mathematics Unit. Indian Statistical Institute, Kolkata, India

Dyckerhoff. R., Mosler, K. & Koshevoy, G. (1996). Zonoid Data Depth: Theory and Computation. In: Prat, A. (Ed.) *COMPSTAT: Proceedings in Computational Statistics 12th Symposium held in Barcelona, Spain, 1996*. Berlin: Physica-Verlag HD. doi: 10.1007/978-3-642-46992-3_26

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179 – 188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Fraiman, R., Meloche, J., García-Escudero, L. A., et al. (1999). Multivariate L-estimation. *Test, 8*(2), pp. 255 – 317. doi: 10.1007/bf02595872

Gao, Y. (2003). Data depth based on spatial rank. *Statistics & Probability Letters, 65*(3), 217 – 225. doi: 10.1016/j.spl.2003.06.003

Ghosh, A. K. & Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics, 32*(2), 327 – 350. doi: 10.1111/j.1467-9469.2005.00423.x

Hall, P., Titterington, D.M. & Xue, J. (2009). Median based classifiers for high dimensional data. *Journal of the American Statistical Association. 104*(488), 1597 – 1608. doi: 10.1198/jasa.2009.tm08107

Hubert, M. & Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis, 45*(2), 301 – 320. doi: 10.1016/s0167-9473(02)00299-2

Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. *The Annals of Statistics, 25*(2), 435 – 477. doi: 10.1214/aos/1031833659

Koshevoy, G. & Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, *25*(5), 1998 – 2017. doi: 10.1214/aos/1069362382

Li, J., Cuesta-Alberstos, J. A. & Liu, R. Y. (2012). DD-classifier: nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association, 107*(498), 737 – 753. doi: 10.1080/01621459.2012.688462

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics, 18*(1), 405 – 414. doi: 10.1214/aos/1176347507

Liu, R. Y. & Singh, K. (1993). A quality index based on multivariate data depth and multivariate rank tests. *Journal of the American Statistical Association, 88*(421), 252 – 260. doi: 10.1080/01621459.1993.10594317

Liu, R. Y., Parelius, J. M. & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics, 27*(3), 783 – 858. doi: 10.1214/aos/1018031260

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*(1), 49–55.

Makinde, O. S. (2015). *On some classification methods for high dimensional and functional data*. PhD Thesis, University of Birmingham

Makinde, O. S. (2017). Multivariate rank outlyingness and correlation effects. *Journal of Modern Applied Statistical Methods*, 16(1), 246-260. doi: 10.22237/jmasm/1493597580

Makinde, O. S. & Chakraborty, B. (2015). On some classifiers based on distribution functions of multivariate ranks. In Nordhausen, K and Taskinen, S. (Eds). *Modern Nonparametric, Robust and Multivariate Methods, Festschrift in Honour of Hannu Oja*. NY: Springer, 249 – 264. doi: 10.1007/978-3-319-22404-6_15

Oja, E. (1983). *Subspace methods of pattern recognition*. Letchworth, England: Research Studies Press.

Olasoji, J. (2014). Inter- and intra-maturity group differences in physiological quality of maize seeds. Unpublished raw data.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871 – 880. doi: 10.1080/01621459.1984.10477105

Rousseeuw, P. J. & Hubert, M. (1999). Regression depth (with discussion). *Journal of the American Statistical Association*, *94*(446), 388 – 433. doi: 10.1080/01621459.1999.10474129

Rousseeuw, P. J. & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212 – 223. doi: 10.1080/00401706.1999.10485670

Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. In Y. Dodge (Ed.). *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pp. 25–38. Basel: Birkhauser. doi: 10.1007/978-3-0348-8201-9_3

Tukey, J. W. (1975). Mathematics and the picturing of data. Proceeding of the International Congress of Mathematicians, Vancouver, 523–531.

Vapnik, V. N. (1998). *Statistical learning theory*. NY: John Wiley and Sons

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*(3-4), 350 – 362. doi: 10.2307/2332010

Zuo, Y. & Serfling, R. (2000a). General notion of statistical depth function. *The Annals of Statistics*, *28*(2), 461 – 482. doi: 10.1214/aos/1016218226

Zuo, Y. & Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth function. *The Annals of Statistics, 28*(2), 483 – 499. doi: 10.1214/aos/1016218227

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 406-427.
doi: 10.22237/jmasm/1493598180

ISSN 1538 − 9472

The Double Prior Selection for the Parameter of Exponential Life Time Model under Type II Censoring

Ronak M. Patel
Som-Lalit College of Commerce
Ahmedabad, India

Achyut C. Patel
Smt. M. T. Dhamsania Community College
Rajkot, India

A comparison of double informative priors assumed for the parameter of exponential life time model is considered. Three different sets of double priors are included, and the results are compared with a forth single prior. The data is Type II censored and Bayes estimators for the parameter and reliability are carried out under a squared error loss function in the cases of the four different sets of prior distributions. The predictive distribution was derived for future failure time and also for the remaining ordered failure times after the first *r* failure times have been observed. Corresponding Bayes credible equal tail intervals are also derived. Simulations and real data are employed to exemplify the method.

Keywords: Gamma prior, chi-square prior, predictive intervals, squared error loss function

Introduction

In life testing experiments, the experimenter may not be always in a position to observe the life times of all items tested because of time limitations or restrictions on the number of failures during the test due to very high-cost items. When the cost of the experiment is directly proportional to the number of failures, the failure-censored experiment is more preferable than the time-censored experiment. The failure-censored experiment is also known as Type II censoring. In this censoring scheme, the test is terminated as soon as the pre-determined number of failures (r) is observed out of n units tested.

The exponential distribution has an important position in life time models. It is the first lifetime model for which statistical methods were extensively developed. Many authors contributed to the methodology of this distribution.

Ronak M. Patel is a Lecturer in the Department of Statistics. Email him at: ronak2307@yahoo.in.

Important works include Epstein (1954), Epstein (1960a, 1960b), Epstein and Sobel (1953), Epstein and Sobel (1954), Bartholomew (1957), Mann, Schafer, and Singpurvala (1974), Lawless (1971), Balakrishnan and Cohen (1991), and others. A number of authors considered prediction problems for the exponential distribution. For example, Hahn (1975), Lawless (1971), and Likeš (1974). In life testing experiments, prediction using the Bayesian method is considered by Box and Tiao (1973), Dunsmore (1974), Evans and Nigm (1980), and Howlader and Hossain (1995). Saleem and Aslam (2008), Tahir and Zawar (2008), and Haq and Dey (2011) considered comparison and selection of a suitable prior using Bayesian methodology.

They considered a set of single prior distributions for comparison. However, there may have been different prior information about the unknown parameter of the lifetime model; to include two different kind of information in the Bayesian analysis, two different priors have been selected for a single unknown parameter of the life time model. Haq and Aslam (2009) considered double prior selection for the parameter of a Poisson distribution based on posterior variance, posterior predictive variance, and the posterior predictive probabilities. Radha and Vekatesan (2013) considered the problem of selection of double priors for the parameter of a Maxwell distribution. They did not derive any Bayes estimator, but just showed that the double priors and their posterior distribution belong to the same family.

The purpose of the current study, therefore, is to contribute something in this direction of double prior distribution for the parameter of the exponential life time model. The following three different types of joint priors and one type of single prior are used for the unknown parameter θ of the exponential distribution:

(i) Exponential-Gamma distribution
(ii) Gamma-Chi-square distribution
(iii) Chi-square-Exponential distribution
(iv) Gamma distribution

Bayes estimators of parameter θ and reliability at time t are obtained based on a Type II censored sample (with fixed r observed failures) under squared error loss function based on the above prior distributions. Also, Bayes predictive estimation and Bayes predictive equal credible intervals are carried out. Prediction of the remaining $(n-r)$ failure times is done and their Bayes credible prediction intervals are also derived. A real-life example is considered to exemplify the

theoretical results obtained in the paper, as is a simulation study, and comparison is made between the results obtained based on the different priors considered.

The Posterior Distribution of θ Under Different Prior Distributions

Let n items be put on a life test such that the test is terminated as soon as the r^{th} failure is observed, and the corresponding failure times are $X_{(1)}$, $X_{(2)}$,…, $X_{(n)}$. During the test, failure items are not replaced, but the test is carried out with the remaining items on the test. Such a censoring scheme is known as Type II censoring without replacement. We assume the life time model is exponential with mean life $1/\theta$, $\theta > 0$. The probability density function (pdf) and the cumulative distribution function (cdf) of this life time model are, respectively, given by

$$\mathrm{f}\left(x,\theta\right)=\theta \mathrm{e}^{-\theta x}, \quad x>0, \theta>0 \tag{1}$$

$$\mathrm{F}\left(x,\theta\right)=1-\mathrm{e}^{-\theta x} \tag{2}$$

Its reliability function at time t is given by

$$\mathrm{R}_{t}\left(\theta\right)=\mathrm{e}^{-\theta t}, \quad t>0 \tag{3}$$

The likelihood function under such a censoring scheme is given by

$$\mathrm{L}\left(x,\theta\right)=C\prod_{i=1}^{r}\mathrm{f}\left(x_{(i)},\theta\right)\left[1-\mathrm{F}\left(x_{(r)},\theta\right)\right]^{n-r}$$

Using (1) and (2),

$$\mathrm{L}\left(x,\theta\right)=C\theta^{r}\mathrm{e}^{-\theta\left[\sum_{i=1}^{r}x_{(i)}+(n-r)x_{(r)}\right]} \tag{4}$$

Exponential and Gamma Distributions as Double Priors

Consider the first prior distribution of θ to be exponential with hyper parameter c_1, having pdf

$$\mathrm{p}_{11}(\theta) = c_1 \mathrm{e}^{-c_1\theta}, \quad \theta > 0, c_1 > 0 \tag{5}$$

The second prior distribution is a gamma distribution with hyper parameters a_1 and b_1, having pdf

$$\mathrm{p}_{12}(\theta) = \frac{\mathrm{e}^{-b_1\theta}\theta^{a_1-1}b_1^{a_1}}{\Gamma(a_1)}, \quad \theta > 0, b_1 > 0, a_1 > 0 \tag{6}$$

The double prior for θ can be defined by combining these two priors as follows:

$$\mathrm{p}_1(\theta) \propto \mathrm{p}_{11}(\theta)\mathrm{p}_{12}(\theta) = k_1 \mathrm{e}^{-(b_1+c_1)\theta}\theta^{a_1-1} \tag{7}$$

where

$$k_1 = \frac{(c_1+b_1)^{a_1}}{\Gamma(a_1)}$$

Hence, the posterior distribution of θ for the given data x is obtained, using (4) and (7), as

$$\pi_1(\theta \mid x) = \frac{(y+c_1+b_1)^{a_1+r}}{\Gamma(a_1+r)} \mathrm{e}^{-(y+b_1+c_1)^{\theta}\theta^{a_1+r-1}}, \quad \theta > 0 \tag{8}$$

where

$$y = \sum_{i=1}^{r} x_{(i)} + (n-r)x_{(r)}$$

which is the gamma distribution with parameters $\alpha_1 = r + a_1$ and $\beta_1 = y + c_1 + b_1$. That is, $\pi_1(\theta \mid x)$ has gamma G(α_1, β_1) distribution.

Gamma and Chi-Square Distributions as Double Priors

Assume the first prior distribution for θ is a gamma distribution:

$$p_{21}(\theta)=\frac{e^{-b_2\theta}\theta^{a_2-1}b_2^{a_2}}{\Gamma(a_2)},\quad \theta>0, b_2>0, a_2>0 \tag{9}$$

The second prior for θ is a chi-square distribution having pdf

$$p_{22}(\theta)=\frac{e^{\frac{-\theta}{2}}\theta^{\left(\frac{c_2}{2}\right)-1}}{2^{\frac{c_2}{2}}\Gamma\left(\frac{c_2}{2}\right)},\quad \theta>0, c_2>0 \tag{10}$$

By combining (9) and (10), obtain the double prior distribution for θ as

$$p_2(\theta)\propto p_{21}(\theta)p_{22}(\theta)=k_2 e^{-\left(b_2+\frac{1}{2}\right)\theta}\theta^{\left(a_2+\frac{c_2}{2}-1\right)-1} \tag{11}$$

where

$$k_2=\frac{\left(b_2+\frac{1}{2}\right)^{a_2+\frac{c_2}{2}-1}}{\Gamma\left(a_2+\frac{c_2}{2}-1\right)}$$

Hence, the posterior distribution of θ based on this double prior distribution of θ for given data x can be obtained, using (4) and (11), as

$$\pi_2(\theta\mid x)=\frac{\left(y+b_2+\frac{1}{2}\right)^{a_2+\frac{c_2}{2}+r-1}}{\Gamma\left(a_2+\frac{c_2}{2}+r-1\right)}e^{-\left(y+b_2+\frac{1}{2}\right)\theta}\theta^{\left(a_2+\frac{c_2}{2}+r-1\right)-1},\quad \theta>0 \tag{12}$$

which is the gamma distribution G(α_2, β_2), where

$$\alpha_2=a_2+\frac{c_2}{2}+r-1,\quad \beta_2=y+b_2+\frac{1}{2}$$

Chi-Square and Exponential Distributions as Double Priors

In a similar manner, assume both the prior distributions have pdfs given by

$$p_{31}(\theta) = \frac{e^{-\frac{\theta}{2}} \theta^{\frac{c_3}{2}-1}}{2^{\frac{c_3}{2}} \Gamma\left(\frac{c_3}{2}\right)}, \quad \theta > 0, c_3 > 0$$

and

$$p_{32}(\theta) = b_3 e^{-b_3 \theta}, \quad \theta > 0, b_3 > 0$$

Hence, the double prior distribution θ becomes

$$p_3(\theta) = k_3 e^{-\left(b_3 + \frac{1}{2}\right)\theta} \theta^{\frac{c_3}{2}-1} \tag{13}$$

where

$$k_3 = \frac{\left(b_3 + \frac{1}{2}\right)^{\frac{c_3}{2}}}{\Gamma\left(\frac{c_3}{2}\right)}$$

and the posterior distribution of θ given the data x, based on this double prior distribution, comes out to be

$$\pi_3(\theta \mid x) = \frac{\left(y + b_3 + \frac{1}{2}\right)^{\frac{c_3}{2}+r}}{\Gamma\left(\frac{c_3}{2} + r\right)} e^{-\left(y + b_3 + \frac{1}{2}\right)\theta} \theta^{\frac{c_3}{2}+r-1}, \quad \theta > 0 \tag{14}$$

which is the gamma distribution $G(\alpha_3, \beta_3)$, where

$$\alpha_3 = \frac{c_3}{2} + r, \quad \beta_3 = y + b_3 + \frac{1}{2}$$

Only Single Gamma G(a_4, b_4) Prior Distribution for θ

Here we consider only a single gamma prior distribution for θ, given by

$$\mathrm{p}_4(\theta) = \frac{\mathrm{e}^{-b_4\theta}\theta^{a_4-1}b_4^{a_4}}{\Gamma(a_4)}, \quad \theta > 0, b_4 > 0, a_4 > 0 \tag{15}$$

The corresponding posterior distribution for θ becomes

$$\pi_4(\theta \mid x) = \frac{(y+b_4)^{a_4+r}}{\Gamma(a_4+r)}\mathrm{e}^{-(y+b_4)\theta}\theta^{a_4+r-1}, \quad \theta > 0 \tag{16}$$

which is also a gamma distribution G(α_4, β_4) with parameters $\alpha_4 = r + a_4$ and $\beta_4 = y + b_4$. Thus, in all the cases of the different types of double prior distributions and in the case of a single prior distribution, the posterior distribution of θ given the data x becomes a gamma distribution. Thus the i^{th} case of the posterior distribution for θ given the data X can be denoted by G(α_i, β_i), i = 1, 2, 3, 4, with pdf

$$\pi_i(\theta \mid x) = \frac{\mathrm{e}^{-\beta_i\theta}\theta^{\alpha_i-1}\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}, \quad \theta > 0, \beta_i > 0, \alpha_i > 0 \tag{17}$$

for i = 1, 2, 3, 4.

Bayes Estimator of θ and Reliability $R_t(\theta)$ at Time t

Consider the squared error loss function defined as

$$\mathrm{L}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

The Bayes estimator of θ under the squared error loss function is nothing but the posterior mean, i.e.

$$\hat{\theta} = \mathrm{E}_{\pi_i}(\theta \mid x) \tag{18}$$

where π_i is the posterior distribution of θ given the data X in case i.

Under the Exponential-Gamma prior distribution, using (8) and (18), the Bayes estimators of θ and $R_t(\theta)$ can be obtained as

$$\begin{aligned}\hat{\theta}_1 &= E_{\pi_1}(\theta \mid x) \\ &= \frac{r + a_1}{y + c_1 + b_1}\end{aligned} \tag{19}$$

with

$$\begin{aligned}V_{\pi_1}(\theta \mid x) &= \frac{\alpha_1}{\beta_1^2} \\ &= \frac{r + a_1}{(y + c_1 + b_1)^2}\end{aligned} \tag{20}$$

and

$$\begin{aligned}\hat{R}_{1t}(\theta) &= \left(\frac{y + c_1 + b_1}{t + y + c_1 + b_1}\right)^{r + a_1} \\ &= \left(\frac{\beta_1}{t + \beta_1}\right)^{\alpha_1}\end{aligned} \tag{21}$$

with

$$\begin{aligned}V_{\pi_1}(R_t(\theta) \mid x) &= \left(\frac{\beta_1}{2t + \beta_1}\right)^{\alpha_1} - \left(\frac{\beta_1}{t + \beta_1}\right)^{2\alpha_1} \\ &= \left(\frac{y + c_1 + b_1}{2t + y + c_1 + b_1}\right)^{r + a_1} - \left(\frac{y + c_1 + b_1}{t + y + c_1 + b_1}\right)^{2(r + a_1)}\end{aligned} \tag{22}$$

Similarly, the Bayes estimators of θ and $R_t(\theta)$ at time t in the case of the i^{th} joint prior distribution are obtained as

$$\hat{\theta}_i = \frac{\alpha_i}{\beta_i} \tag{23}$$

with

$$V_{\pi_i}(\theta \mid x) = \frac{\alpha_i}{\beta_i^2} \tag{24}$$

and

$$\hat{R}_{it}(\theta) = \left(\frac{\beta_i}{t+\beta_i}\right)^{\alpha_i} \tag{25}$$

with

$$V_{\pi_i}(R_t(\theta) \mid x) = \left(\frac{\beta_i}{2t+\beta_i}\right)^{\alpha_i} - \left(\frac{\beta_i}{t+\beta_i}\right)^{2\alpha_i} \tag{26}$$

for $i = 2, 3, 4$.

$(1 - \alpha)100\%$ Equal Tail Credible Interval for θ

Let $[I_{1i}, I_{2i}]$ be the $(1 - \alpha)100\%$ equal tail credible interval for θ. Then I_{1i} and I_{2i} can be obtained by solving the following equations:

$$\int_0^{I_{1i}} \pi_i(\theta \mid x)\, d\theta = \frac{\alpha}{2} \quad \text{and} \quad \int_{I_{2i}}^{\infty} \pi_i(\theta \mid x)\, d\theta = \frac{\alpha}{2} \tag{27}$$

From (17) and (27),

$$G(\alpha_i, \beta_i, I_{1i}) = \frac{\alpha}{2} \quad \text{and} \quad G(\alpha_i, \beta_i, I_{2i}) = 1 - \frac{\alpha}{2}$$

where

$$\begin{aligned} G(n, a, I) &= \int_0^I \frac{e^{-a\theta}\theta^{n-1}a^n}{\Gamma(a)}\, d\theta \\ &= \sum_{j=0}^{n-1} \frac{e^{-Ia}(Ia)^j}{j!} \end{aligned} \tag{28}$$

Thus, equation (27) reduces to

$$\sum_{j=0}^{\alpha_i-1}\frac{e^{-I_{1i}\beta_i}\left(I_{1i}\beta_i\right)^j}{j!}=\frac{\alpha}{2}\quad\text{and}\quad\sum_{j=0}^{\alpha_i-1}\frac{e^{-I_{2i}\beta_i}\left(I_{2i}\beta_i\right)^j}{j!}=1-\frac{\alpha}{2}\tag{29}$$

Solving these equations yields I_{1i} and I_{2i}, $i = 1, 2, 3, 4$.

Posterior Distribution of $R_t(\theta)$

As $R_t(\theta) = e^{-t\theta}$, from the posterior distribution of θ given the data x as defined in (17), the posterior distribution of $R_t(\theta) = R$ can be derived as

$$\pi_i\left(R\mid x\right)=\left(\frac{\beta_i}{t}\right)^{\alpha_i}\frac{1}{\Gamma\left(\alpha_i\right)}\left(R\right)^{\frac{\beta_i}{t}-1}\left(-\ln R\right)^{\alpha_i-1},\quad 0<R<1\tag{30}$$

The $(1-\alpha)100\%$ equal tail credible interval for $R(t)$ can be derived by solving the equations:

$$\int_0^{R_{1i}}\pi_i\left(R\mid x\right)d\theta=\frac{\alpha}{2}\quad\text{and}\quad\int_{R_{2i}}^{\infty}\pi_i\left(R\mid x\right)d\theta=\frac{\alpha}{2}\tag{31}$$

From (30) and (31),

$$G\left(\alpha_i,\frac{\beta_i}{t},-\ln R_{1i}\right)=\frac{\alpha}{2}\quad\text{and}\quad G\left(\alpha_i,\frac{\beta_i}{t},-\ln R_{2i}\right)=1-\frac{\alpha}{2}$$

Again using (28), deduce that

$$\sum_{j=0}^{\alpha_i-1}\frac{\left(R_{1i}\right)^{\frac{\beta_i}{t}}\left(-\ln R_{1i}\frac{\beta_i}{t}\right)^j}{j!}=\frac{\alpha}{2}\quad\text{and}\quad\sum_{j=0}^{\alpha_i-1}\frac{\left(R_{2i}\right)^{\frac{\beta_i}{t}}\left(-\ln R_{2i}\frac{\beta_i}{t}\right)^j}{j!}=1-\frac{\alpha}{2}\tag{32}$$

Finally, solving these equations, obtain R_{1i} and R_{2i} for $i = 1, 2, 3, 4$.

Bayes Predictive Estimator and Equal Tail Credible Interval for a Future Observation

A predictive estimator is derived for a future observation and its equal tail credible interval. Let Z_i be a future observation which has already survived $X_{(r)}$, and let $W_{(i)} = Z_i - X_{(r)}$. Given the data x, the conditional joint pdf of W_i and θ is

$$\mathrm{h}_i\left(w_i, \theta \mid x\right) = \mathrm{f}\left(w_i, \theta \mid x\right) \pi_i\left(\theta \mid x\right)$$

From (10) and (17),

$$\mathrm{h}_i\left(w_i, \theta \mid x\right) = \frac{\beta_i^{\alpha_i}}{\Gamma\left(\alpha_i\right)} \mathrm{e}^{-\theta\left(w_i+\beta_i\right)} \theta^{\alpha_i} \tag{33}$$

Integrating out with respect to θ, the predictive density of w_i under the i^{th} case of the joint prior distribution comes out as

$$\begin{aligned} \mathrm{p}_i\left(w_i \mid x\right) &= \int_0^{\infty} \frac{\beta_i^{\alpha_i}}{\Gamma\left(\alpha_i\right)} \mathrm{e}^{-\theta\left(w_i+\beta_i\right)} \theta^{\alpha_i} d\theta \\ &= \frac{\alpha_i \beta_i^{\alpha_i}}{\left(w_i+\beta_i\right)^{\alpha_i+1}}, \quad w_i > 0, \alpha_i > 0, \beta_i > 0, i = 1,2,3,4 \end{aligned} \tag{34}$$

The Bayes estimator of w_i under a squared error loss function is given by

$$\begin{aligned} w_i^* &= \mathrm{E}\left(w_i \mid x\right) = \alpha_i \beta_i^{\alpha_i} \int_0^{\infty} \frac{w_i}{\left(w_i+\beta_i\right)^{\alpha_i+1}} dw_i \\ &= \frac{\beta_i}{\alpha_i - 1}, \quad i = 1,2,3,4 \end{aligned} \tag{35}$$

Hence

$$z_i^* = x_{(r)} + w_i^* \tag{36}$$

Now the $(1-\alpha)100\%$ predictive interval (h_{1i}, h_{2i}) for w_i can be obtained by solving the equations

$$\int_0^{h_{1i}} \mathrm{p}_i\left(w_i \mid x\right) dw_i = \frac{\alpha}{2} \quad \text{and} \quad \int_0^{h_{2i}} \mathrm{p}_i\left(w_i \mid x\right) dw_i = 1 - \frac{\alpha}{2} \tag{37}$$

Using (34) in (37), after some algebraic manipulation,

$$h_{1i} = \beta_i \frac{1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\alpha_i}}}{\left(\frac{\alpha}{2}\right)^{\frac{1}{\alpha_i}}} \quad \text{and} \quad h_{2i} = \beta_i \frac{1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\alpha_i}}}{\left(\frac{\alpha}{2}\right)^{\frac{1}{\alpha_i}}}, \quad i = 1, 2, 3, 4 \tag{38}$$

Prediction of the Remaining ($n - r$) Failure Times Truncated at $x_{(r)}$ and their Equal Tail Intervals

Consider the prediction of the remaining $(n - r)$ failure times given the first r failure times of a sample of n units. Let $x_{(s)i}$, $(r + 1 \le s \le n)$ denote the failure times of the s^{th} unit to fail under the i^{th} case of double prior distribution. The conditional pdf of $u = x_{(s)} - x_{(r)}$ from a pdf truncated at $x_{(r)}$ is given by

$$\mathrm{f}\left(u \mid \theta\right) = \frac{\left(\mathrm{f}\left(u\right)\right)^{s-r-1} \left[1 - \mathrm{F}\left(u\right)\right]^{n-s} \mathrm{f}\left(u\right)}{\beta_{(s-r, n-s+1)}}, \quad u \ge 0$$

Here f(u) and F(u) are the pdf and cdf of a random variable X, respectively, as given in (1) and (2). Hence

$$\mathrm{f}\left(u \mid \theta\right) = \frac{\left(1 - \mathrm{e}^{-\theta u}\right)^{s-r-1} \left[\mathrm{e}^{-\theta u}\right]^{n-s} \theta \mathrm{e}^{-\theta u}}{\beta_{(s-r, n-s+1)}}, \quad u \ge 0, r + 1 \le s \le n \tag{39}$$

Given x, the conditional joint pdf of u and θ under the i^{th} case of the double prior distribution is given by

$$\begin{aligned} \mathrm{f}_i\left(u_i,\theta \mid x\right) &= \mathrm{f}\left(u_i \mid \theta\right)\pi_i\left(\theta \mid x\right) \\ &= \frac{\left(1-\mathrm{e}^{-\theta u_i}\right)^{s-r-1}\left[\mathrm{e}^{-\theta u_i}\right]^{n-s}\theta \mathrm{e}^{-\theta u_i}}{\beta_{(s-r,n-s+1)}}\frac{\beta_i^{\alpha_i}}{\Gamma\left(\alpha_i\right)}\mathrm{e}^{-\theta\beta_i}\theta^{\alpha_i-1} \\ &= \frac{\beta_i^{\alpha_i}}{\Gamma\left(\alpha_i\right)\beta_{(s-r,n-s+1)}}\sum_{j=0}^{s-r-1}\frac{(-1)^j\left(s-r-1\right)!}{j!\left(s-r-j-1\right)!}\mathrm{e}^{-\theta\left(ju_i+u_i\left(n-s+1\right)+\beta_i\right)}\theta^{\alpha_i} \end{aligned} \tag{40}$$

Integrating out θ, the predictive density of u_i for the i^{th} case of the double prior is given by

$$\mathrm{p}_i\left(u_i \mid x\right) = \frac{\beta_i^{\alpha_i}}{\Gamma\left(\alpha_i\right)\beta_{(s-r,n-s+1)}}\sum_{j=0}^{s-r-1}\frac{(-1)^j\left(s-r-1\right)!}{j!\left(s-r-j-1\right)!}\frac{\Gamma\left(\alpha_i+1\right)}{\left[ju_i+u_i\left(n-s+1\right)+\beta_i\right]^{\alpha_i+1}} \tag{41}$$

with $u_i > 0$. Under the squared error loss function, the Bayes predictive estimator of u_i is

$$\begin{aligned} u_i^* &= \mathrm{E}\left(u_i \mid x\right) \\ &= \int_0^\infty u_i\,\mathrm{p}_i\left(u_i \mid x\right)du_i \end{aligned}$$

On simplification,

$$u_i^* = \frac{\alpha_i\beta_i}{\beta_{(s-r,n-s+1)}}\sum_{j=0}^{s-r-1}\frac{(-1)^j\left(s-r-1\right)!\beta_{(2,\alpha_i-1)}}{j!\left(s-r-j-1\right)!\left(j+n-s+1\right)^2}, \quad i=1,2,3,4 \tag{42}$$

For $s = r + 1$, obtain the estimator for the $x_{(r+1)}{}^{\text{th}}$ failure time as $x_{(r+1)} = u_i^* + x_{(r)}$, where

$$u_i^* = \frac{\beta_i}{\left(\alpha_i-1\right)\left(n-r\right)}, \quad i=1,2,3,4 \tag{43}$$

Similarly, for $s = r + 2, r + 3,\ldots, n$, u_i^* can be obtained. Now a $(1 - \alpha)100\%$ equal tail confidence interval (H_{1i}, H_{2i}) for u_i is the solution of the equations:

$$\int_0^{H_{1i}} \mathrm{p}_i\left(u_i \mid x\right) du_i = \frac{\alpha}{2} \quad \text{and} \quad \int_0^{H_{2i}} \mathrm{p}_i\left(u_i \mid x\right) du_i = 1 - \frac{\alpha}{2} \tag{44}$$

Using (41) and (44), we have

$$\frac{1}{\beta_{(s-r,n-s+1)}} \sum_{j=0}^{s-r-1} \frac{(-1)^j (s-r-1)!}{j!(s-r-j-1)!} \left[1 - \frac{\beta_i^{\alpha_i}}{\left\{H_{1i}\left(j+n-s+1\right)+\beta_i\right\}^{\alpha_i}}\right] = \frac{\alpha}{2}$$

and (45)

$$\frac{1}{\beta_{(s-r,n-s+1)}} \sum_{j=0}^{s-r-1} \frac{(-1)^j (s-r-1)!}{j!(s-r-j-1)!} \left[1 - \frac{\beta_i^{\alpha_i}}{\left\{H_{2i}\left(j+n-s+1\right)+\beta_i\right\}^{\alpha_i}}\right] = 1 - \frac{\alpha}{2}$$

Solving the equations in (45), obtain the $(1-\alpha)100\%$ equal tail confidence interval (H_{1i}, H_{2i}) for the remaining $(n-r)$ failure times given the first r failure times of a sample of size n for the i^{th} case of the double prior distribution, $i = 1, 2, 3, 4$.

A Real Data Example

The data were obtained from Bain and Engelhardt (1991), representing the times between successive failures. The times are exponentially distributed (Kolmogorov-Smirnov p-value: 0.900) with mean failure time 3.744.

Times between system failures data:

5.2, 8.4, 0.9, 0.1, 5.9, 17.9, 3.6, 2.5, 1.2, 1.8, 1.8, 6.1, 5.3, 1.2, 1.2, 3.0, 3.5, 7.6, 3.4, 0.5, 2.4, 5.3, 1.9, 2.8, 0.1

For this example, Bayes estimates of parameter θ, reliability R(t), Bayes predictive estimator of a future observation Z^*, and predictive estimator of the remaining order statistic $x_{(r+1)}$ based on the known first r order statistics and their equal tail credible intervals (in the bracket) are derived under different types of joint prior distributions and presented in Table 1 for hyper parametric values $a_i = b_i = c_i = 4$, $i = 1, 2, 3, 4$.

Table 1. Bayes estimates and credible intervals

Prior Distribution	$\hat{\theta}$	$\hat{R}(t)$	Z^*	$X_{(r+1)}$
Exponential-Gamma	0.291971	0.242397	8.873913	6.014783
	(0.187679,	(0.122587,	(5.386759,	(5.303434,
	0.419851)	0.92453)	18.95711)	9.361143)
Gamma-Chi-square	0.317662	0.214404	8.579167	5.955853
	(0.205576,	(0.103450,	(5.379741,	(5.303156,
	0.453757)	0.357776)	17.81308)	9.015658)
Chi-square-Exponential	0.279543	0.257919	9.047619	6.049524
	(0.175190,	(0.130106,	(5.390621,	(5.303587,
	0.407894)	0.416476)	19.66698)	9.586129)
Only Gamma	0.306905	0.225945	8.700000	5.980000
	(0.196647,	(0.110079,	(5.382537,	(5.303267,
	0.441330)	0.374119)	18.29253)	9.16352)

Simulation Study

A Monte Carlo simulation study was carried out to compare the performance of the Bayes estimators under different joint priors and single prior. To generate 1000 Type II censored samples, the value of the parameter θ is considered as 0.7 and the values of the hyper parameters for all joint and single priors are considered to be $a_i = b_i = c_i = 4$. The reliability is calculated at time $t = 1$.

Values of the hyper parameters can be obtained from our prior belief. If there is any information from past data about the mean, variance, or about reliability measure, by comparing such prior beliefs with the theoretical results and by solving the equations the estimates of the hyper parameters can be obtained. A value 4 was used for the hyper parameters, i.e. $a_i = b_i = c_i = 4$, for simulation purpose.

The simulation was conducted for different values of sample size (n) and of fixed censored value (r): (n, r) = (15, 5), (15, 10), (15, 12), (30, 10), (30, 20), and (30, 24). In each case, Bayes estimates of θ, R(t), future observation z^*, and the $(r + 1)^{th}$ ordered failure time $X_{(r+1)}$ are derived. Their mean square errors (MSE) and Bayes equal tail credible intervals are also obtained. The first, second, and third values for each cell of the third and fourth columns of Tables 2 to 5 denote the Bayes estimate, MSE, and credible intervals, respectively.

Table 2. Bayes estimates and credible intervals for θ and R(t) for n = 15

Joint priors	r	θ	R(t)
Exponential-Gamma	5	0.620546	0.552537
		0.020495	0.006874
		(0.283758, 1.08687)	(0.344604, 0.754072)
	10	0.658255	0.529796
		0.018407	0.005175
		(0.359880, 1.045241)	(0.358859, 0.699491)
	12	0.666924	0.524325
		0.017280	0.004657
		0.381210, 1.031245)	(0.363315, 0.684830)
Gamma-Chi-square	5	0.923133	0.423851
		0.105171	0.012966
		(0.442683, 1.577159)	(0.223643, 0.646348)
	10	0.852047	0.444890
		0.064292	0.009208
		0.476877, 1.334254)	(0.276191, 0.624641)
	12	0.835802	0.449584
		0.054266	0.008109
		(0.486891, 1.277455)	(0.289965, 0.618192)
Chi-square-Exponential	5	0.646193	0.545833
		0.030033	0.008780
		(0.259808, 1.205569)	(0.313306, 0.772884)
	10	0.681619	0.521700
		0.026693	0.006481
		(0.352208, 1.117977)	(0.338067, 0.705574)
	12	0.688308	0.516572
		0.024432	0.005762
		(0.376310, 1.092958)	(0.345085, 0.688844)
Only Gamma	5	0.873913	0.444968
		0.085159	0.010865
		(0.399614, 1.530634)	(0.234776, 0.674373)
	10	0.819872	0.459206
		0.055039	0.008196
		(0.448236, 1.301867)	(0.285404, 0.642552)
	12	0.807536	0.462281
		0.046925	0.007281
		(0.461582, 1.248669)	(0.298505, 0.633848)

Table 3. Bayes estimates and credible intervals for θ and R(t) for n = 30

Joint priors	r	θ	R(t)
Exponential-Gamma	10	0.658399	0.529698
		0.018254	0.005154
		(0.359958, 1.045468)	(0.358739, 0.699423)
	20	0.681753	0.513881
		0.013730	0.003502
		(0.436817, 0.980345)	(0.380295, 0.647858)
	24	0.686436	0.510669
		0.012823	0.003193
		(0.456137, 0.963065)	(0.386387, 0.635486)
Gamma-Chi-square	10	0.852159	0.444759
		0.063765	0.009202
		(0.476952, 1.334466)	(0.275997, 0.624550)
	20	0.791372	0.463756
		0.030889	0.005325
		(0.512140, 1.130405)	(0.330128, 0.602004)
	24	0.779839	0.467613
		0.026108	0.004659
		(0.522275, 1.088224)	(0.343117, 0.595763)
Chi-square-Exponential	10	0.681728	0.521585
		0.026326	0.006450
		(0.352264, 1.118155)	(0.337909, 0.705506)
	20	0.696408	0.508011
		0.017468	0.004104
		(0.436440, 1.016152)	(0.368543, 0.648527)
	24	0.699166	0.505450
		0.015863	0.003687
		(0.456723, 0.992416)	(0.376460, 0.635478)
Only Gamma	10	0.819986	0.459075
		0.054456	0.008184
		(0.448299, 1.302049)	(0.285213, 0.642467)
	20	0.772391	0.472523
		0.027461	0.004911
		(0.494890, 1.110679)	(0.336687, 0.612383)
	24	0.763553	0.475196
		0.023514	0.004341
		(0.507380, 1.071258)	(0.348968, 0.604628)

Table 4. Bayes estimates and credible intervals for Z^* and $x_{(r+1)}$ for $n = 15$

Joint priors	r	Z^*	$x_{(r+1)}$
Exponential-Gamma	5	2.435166	0.738994
		0.394374	0.078934
		(0.593002, 8.189055)	(0.550949, 1.976656)
	10	3.161177	1.800628
		0.670290	0.303256
		(1.50051, 8.125572)	(1.462075, 3.494605)
	12	3.738878	2.633385
		0.880619	0.541882
		(2.120031, 8.530325)	(2.084978, 4.972641)
Gamma-Chi-square	5	1.836860	0.679165 0.076552
		0.342344	0.076552
		(0.579878, 5.715377)	(0.550820, 1.500505)
	10	2.789699	1.726333
		0.631423	0.298065
		(1.491927, 6.648723)	(1.461736, 3.037210)
	12	3.416487	2.525922
		0.845020	0.532755
		(2.112493, 7.260214)	(2.084148, 4.398042)
Chi-square-Exponential	5	2.480049	0.743482
		0.572588	0.086304
		(0.592479, 83.582894)	(0.5506944, 2.117547)
	10	3.152210	1.798834
		0.774620	0.316687
		(1.499793, 8.157714)	(1.462047, 3.526374)
	12	3.724764	2.628682
		0.971637	0.564728
		(2.119327, 8.524054)	(2.084899, 4.985378)
Only Gamma	5	1.935169	0.688994
		0.394375	0.078935
		(0.581735, 6.162530)	(0.55084, 1.598301)
	10	2.853482	1.739090
		0.670290	0.303256
		(1.493269, 6.919713)	(1.461789, 3.1265887)
	12	3.472212	2.544495
		0.880620	0.548882
		(2.113695, 7.493128)	(2.084279, 4.507569)

Table 5. Bayes estimates and credible intervals for Z^* and $x_{(r+1)}$ for $n = 30$

Joint priors	r	Z^*	$x_{(r+1)}$
Exponential-Gamma	10	2.259726	0.644702
		0.268444	0.037871
		(0.599704, 7.222201)	(0.559800, 1.235965)
	20	3.081523	1.663868
		0.377836	0.142476
		(1.544589, 7.525605)	(1.506728, 2.533688)
	24	3.735099	2.442472
		0.521211	0.273523
		(2.221829, 8.081593)	(2.184982, 3.693112)
Gamma-Chi-square	10	1.888297	0.626132
		0.243800	0.037402
		(0.591122, 5.745548)	(0.559779, 1.081897)
	20	2.870057	1.642720
		0.364330	0.141661
		(1.53951, 6.710169)	(1.506678, 2.392701)
	24	3.554705	2.412403
		0.508779	0.272072
		(2.217467, 7.390260)	(2.184864, 3.514660)
Chi-square-Exponential	10	2.250640	0.644248
		0.336536	0.039082
		(0.598986, 7.253850)	(0.559799, 1.253045)
	20	3.064872	1.662202
		0.409659	0.144348
		(1.544036, 7.481150)	(1.506722, 2.530874)
	24	3.719194	2.439816
		0.549649	0.276788
		(2.221333, 8.034713)	(2.184969, 3.685021)
Only Gamma	10	1.952034	0.629318
		0.268442	0.037872
		(0.592464, 6.016336)	(0.559782, 1.113566)
	20	2.907609	1.646474
		0.377836	0.142476
		(1.540366, 6.861025)	(1.506685, 2.420264)
	24	3.586950	2.417776
		0.521211	0.273523
		(2.218210, 7.518306)	(2.184883, 3.548975)

Conclusion

Comparison of Priors Based on the MSE and Credible Interval of θ

From the third columns of Tables 2 and 3 it is observed that the values of the MSE of the Bayes estimator of parameter θ and length of its credible intervals are smaller in the case of the Exponential-Gamma joint prior and then followed by Chi-square-Exponential, Gamma, and Gamma-Chi-square priors in all the values of n and r considered here.

Comparison of Priors Based on the MSE and Credible Interval of R(t)

From the fourth column of Tables 2 and 3 it is observed that, for all values of n and r considered here, the values of the MSE of the Bayes estimator of R(t) are smaller in the case of the joint prior Exponential-Gamma, and then followed by Chi-square-Exponential, Gamma, and Gamma-Chi-square priors. The Exponential-Gamma joint prior generates the minimum length of the credible intervals of R(t), followed by Gamma-Chi-square, Gamma, and Chi-square-Exponential priors in all the values of n and r considered here.

Comparison of Priors Based on MSE and Credible Interval of Future Predicted Value

From the third columns of Tables 4 and 5, note that MSEs and lengths of the credible intervals are minimum in the case of the Gamma-Chi-square prior, and then followed by Gamma, Exponential-Gamma, and Chi-square-Exponential priors in all the values of n and r considered here.

Comparison Based on the MSE and Credible Interval of Next Ordered Failure Time $X_{(r+1)}$

From the fourth columns of the Tables 4 and 5, we observed that the minimum MSE as well as the minimum length of credible interval are generated by the Gamma-Chi-square joint prior, whereas other priors give erratic effect.

Thus, it was found that the Exponential-Gamma joint prior performs well compared to the other single and joint priors considered in this study.

Acknowledgements

The authors are thankful to the honorable referees and to the editor of the journal for making good comments on the paper.

References

Bain, L. J., & Engelhardt, M. (1991). *Statistical analysis of reliability and life testing models*. New York, NY: Marcel Dekker.

Balakrishnan, N., & Cohen, A. C. (1991). *Order statistics and inference: Estimation methods*. Boston, MA: Academic Press.

Bartholomew, D. J. (1957). A problem in life testing. *Journal of the American Statistical Association, 52*(279), 350-355. doi: 10.2307/2280905

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addition-Wesley.

Dunsmore, I. R. (1974). The Baysian predictive distribution in life testing. *Technometrics, 16*(3), 455-466. doi: 10.1080/00401706.1974.10489216

Epstein, B. (1954). Truncated life tests in the exponential case. *The Annals of Mathematical Statistics, 25*(3), 555-564. doi: 10.1214/aoms/1177728723

Epstein, B. (1960a). Tests for the validity of the assumption that the underlying distribution of life is exponential. Part I. *Technometrics, 2*(1), 83-101. doi: 10.2307/1266533

Epstein, B. (1960b). Tests for the validity of the assumption that the underlying distribution of life is exponential. Part II. *Technometrics, 2*(1), 167-183. doi: 10.2307/1266543

Epstein, B., & Sobel, M. (1953). Life testing. *Journal of the American Statistical Association, 48*(263), 486-502. doi: 10.2307/2281004

Epstein, B., & Sobel, M. (1954). Some problems relevant to life testing from an exponential distribution. *The Annals of Mathematical Statistics, 25*(2), 373-381. doi: 10.1214/aoms/1177728793

Evans, I. G., & Nigm, A. H. M. (1980). Bayesian prediction for the left truncated exponential distribution. *Technometrics, 22*(2), 201-204. doi: 10.2307/1268459

Hahn, G. J. (1975). A Simultaneous prediction limit on the mean of future samples from an exponential distribution. *Technometrics, 17*(3), 341-345. doi: 10.2307/1268071

Haq, A., & Aslam, M. (2009). On the double prior selection for the parameter of the parameter of Poisson distribution. *InterStat: Statistics on the Internet, 2009*(November), #1. Retrieved from http://interstat.statjournals.net/YEAR/2009/articles/0911001.pdf

Haq, A., & Dey, S. (2011). Bayesian estimation of Erlang distribution under different prior distributions. *Journal of Reliability and Statistical Studies, 4*(1), 1-30.

Howlader, H. A., & Hossain, A. (1995). On Bayesian estimation and prediction from Rayleigh based on Type II censored data. *Communications in Statistics – Theory and Methods, 24*(9), 2249-2259. doi: 10.1080/03610929508831614

Lawless, J. F. (1971). A prediction problem concerning samples from the exponential distribution, with applications in life testing. *Technometrics, 13*(4), 725-730. doi: 10.2307/1266949

Likeš, J. (1974). Prediction of sth ordered observations for the two-parameter exponential distribution. *Technometrics, 16*(2), 241-244. doi: 10.2307/1267945

Mann, N. R., Schafer, R. E., & Singpurwalla, N. D. (1974). *Methods for statistical analysis of reliability and life data*. New York, NY: John Wiley and Sons.

Radha, R. K., & Vekatesan, P. V. (2013). On the double prior selection for the parameter of Maxwell distribution. *International Journal of Scientific & Engineering Research, 4*(5), 1238-1241. Retrieved from http://www.ijser.org/researchpaper%5CON-THE-DOUBLE-PRIOR-SELECTION-FOR-THE-PARAMETER-OF-MAXWELL-DISTRIBUTION.pdf

Saleem, M., & Aslam, M. (2008). On the prior selection for the mixture of Rayleigh distribution using predictive intervals. *Pakistan Journal of Statistics, 24*(1), 21-35.

Tahir, M., & Zawar, H. (2008). Comparison of non-informative priors for number of defect (Poisson) model. *InterStat: Statistics on the Internet, 2008*(April), #2. Retrieved from http://interstat.statjournals.net/YEAR/2008/articles/0804002.pdf

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 428-451.
doi: 10.22237/jmasm/1493598240

ISSN 1538 − 9472

Monte Carlo Study of Some Classification-Based Ridge Parameter Estimators

A. F. Lukman
Ladoke Akintola Univ. of Technology
Ogbomosho, Nigeria

K. Ayinde
Ladoke Akintola Univ. of Technology
Ogbomosho, Nigeria

A. S. Ajiboye
Federal University of Technology
Akure, Nigeria

Ridge estimator in linear regression model requires a ridge parameter, K, of which many have been proposed. In this study, estimators based on Dorugade (2014) and Adnan et al. (2014) were classified into different forms and various types using the idea of Lukman and Ayinde (2015). Some new ridge estimators were proposed. Results shows that the proposed estimators based on Adnan et al. (2014) perform generally better than the existing ones.

Keywords: linear regression model, multicollinearity, ridge estimator, mean square error

Introduction

The parameter estimates obtained through the use of the Ordinary Least Squares (OLS) estimator have optimal performance when there is no violation of any of the assumptions of the classical linear regression model. One of the most basic of these assumptions is that explanatory variables are independent. Multicollinearity refers to the presence of strong or perfect linear relationships among the explanatory variables. Multicollinearity is an inherent phenomenon in most economic relationships due to the nature of economic magnitude (Koutsoyiannis, 2003). When there is a perfect relationship among the explanatory variables, the regression coefficients of the OLS estimator are indeterminate, and the standard error of the estimates becomes very large. Also, when there are strong relationships among the explanatory variables, the regression estimates are determinate but possesses large standard error (Koutsoyiannis, 2003).

Adewale Folaranmi Lukman is a teaching assistant in the Department of Statistics. Email at wale3005@yahoo.com. Prof. Kayode Ayinde is a lecturer in the Department of Statistics. Email at kayinde@lautech.edu.ng. Dr. Ajiboye S. Adegoke is a lecturer in the Department of Statistics.

Generally, the performance of OLS estimator is unsatisfactory when there is multicollinearity (Koutsoyiannis, 2003). Several techniques have been suggested in the literature to handle this problem. Massy (1965) introduced the principal component regression to eliminate the model instability and reduce the variances of the regression coefficients. Wold (1966) developed the partial least square to deal with the problem of multicollinearity. Hoerl and Kennard (1970) proposed the ridge estimator for dealing with multicollinearity in a regression model, which modifies the OLS to allow biased estimation of the regression coefficients. This study is limited to the application of the ridge regression estimator in handling the problem of multicollinearity. Ridge estimator is defined as:

$$\hat{\beta}_R = \left(X'X + \mathrm{KI} \right)^{-1} X'Y \tag{1}$$

where K is a non-negative constant known as ridge parameter and I denotes an identity matrix. When K equals zero, (1) returns to OLS estimator; this is defined as follows:

$$\hat{\beta}_{OLS} = \left(X'X \right)^{-1} X'Y \tag{2}$$

The corresponding mean square error (MSE) of (1) and (2) are defined respectively as:

$$MSE\left(\hat{\beta}_R \right) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{\lambda_i}{\left(\lambda_i + \hat{K} \right)^2} + \hat{K}^2 \sum_{i=1}^{p} \frac{\hat{\beta}_i^2}{\left(\lambda_i + \hat{K} \right)^2} \tag{3}$$

$$MSE\left(\hat{\beta}_{OLS} \right) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{4}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $X'X$, $\hat{K}$ is the estimator of the ridge parameter K and $\hat{\beta}_i$ is the i^{th} element of the vector $\hat{\beta}$.

Although this estimator is biased, it gives a smaller mean squared error when compared to the OLS estimator for a positive value of K (Hoerl and Kennard, 1970). The use of the estimator depends largely on the ridge parameter, K. Several methods for estimating this ridge parameter have been proposed by different authors, as follows: Hoerl and Kennard (1970); McDonald and

Galarneau (1975); Lawless and Wang (1976); Hocking et al. (1976); Wichern and Churchill (1978); Gibbons (1981); Nordberg (1982); Kibria (2003), Khalaf and Shukur (2005), Alkhamisi et al. (2006), Muniz and Kibria (2009), Mansson et al. (2010), Dorugade (2014) and recently, Lukman and Ayinde (2015). The purpose of this study is to classify the ridge parameters proposed by Dorugade (2014) and Adnan et al. (2014) into different forms and various types. A simulation study is conducted and the performances of the estimators is examined via mean square error (MSE).

Model and Estimators

A linear regression model can be expressed in matrix form as:

$$Y = \mathrm{X}\beta + \mathrm{U} \tag{5}$$

where X is an $n \times p$ matrix with full rank, Y is a $n \times 1$ vector of dependent variable, β is a $p \times 1$ vector of unknown parameters, and U is the error term such that $\mathrm{E(U)} = 0$ and $\mathrm{E(UU')} = \sigma^2 \mathrm{I}_n$. The Ordinary Least Square (OLS) estimator of β is defined in (2): Model (5) can be written in canonical form. Suppose there exists an orthogonal matrix Q such that $X'\mathrm{Q}X = \Lambda$, where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $X'X$. Substituting $\alpha = \mathrm{Q}'\beta$, model (5) can be written as:

$$Y = \mathrm{Z}\alpha + \mathrm{U} \tag{6}$$

where $Z'\mathrm{Z} = \Lambda$.

Therefore, the ridge estimator of α can be defined as:

$$\hat{\alpha}_R = \left(Z'\mathrm{Z} + \mathrm{KI}\right)^{-1} Z'Y \tag{7}$$

The corresponding mean square error (MSE) is defined as:

$$MSE\left(\hat{\alpha}_R\right) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{\lambda_i}{\left(\lambda_i + \hat{\mathrm{K}}\right)^2} + \hat{\mathrm{K}}^2 \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\left(\lambda_i + \hat{\mathrm{K}}\right)^2} \tag{8}$$

where $\hat{\alpha}_i$ is the i^{th} element of the vector $\alpha = Q'\beta$. Hoerl and Kennard (1970) defined the value of the ridge parameter K that minimizes the mean square error as:

$$\hat{\mathrm{K}}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}, \text{ where } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_1^2}{n-p} \tag{9}$$

Hoerl and Kennard (1970) proposed

$$\hat{\mathrm{K}}_{\mathrm{HK}i} = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}.$$

They suggested estimating ridge parameter by taking the maximum (Fixed Maximum) of α_i^2 such that the estimator of K is:

$$\hat{\mathrm{k}}_{\mathrm{HK}}^{\mathrm{FM}} = \frac{\hat{\sigma}^2}{\max\left(\hat{\alpha}_i^2\right)} \tag{10}$$

Hoerl et al. (1975) proposed a different estimator of K by taking the Harmonic Mean of the ridge parameter $\mathrm{K}_{\mathrm{HK}i}$. This estimator is given as:

$$\hat{\mathrm{K}}_{\mathrm{HK}}^{\mathrm{HM}} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p} \alpha_i^2} \tag{11}$$

Kibria (2003) proposed some new estimators of K by taking the geometric mean, arithmetic mean and median ($p \geq 3$) of the ridge parameter $\mathrm{K}_{\mathrm{HK}i}$. These estimators are respectively defined as:

$$\hat{\mathrm{K}}_{\mathrm{HK}}^{\mathrm{GM}} = \frac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p} \hat{\alpha}_i^2\right)^{\frac{1}{p}}} \tag{12}$$

$$\hat{\mathrm{K}}_{\mathrm{HK}}^{\mathrm{AM}} = \frac{\hat{\sigma}^2}{p} \sum_{i=1}^{p} \frac{1}{\hat{\alpha}_i^2} \tag{13}$$

$$\hat{K}_{HK}^{M} = \text{Median}\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\right) \tag{14}$$

Furthermore, Muniz and Kibria (2009) proposed some estimators of K in the form of the square root of the geometric mean of K_{HK_i} and its reciprocal, the median of the square root of K_{HK_i} and its reciprocal, and varying maximum of the square root of K_{HK_i} and its reciprocal. These estimators are respectively defined as:

$$\hat{K}_{HK}^{GMSR} = \sqrt{\frac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p}\hat{\alpha}_i^2\right)^{\frac{1}{p}}}} \tag{15}$$

$$\hat{K}_{HK}^{GMRSR} = \frac{1}{\sqrt{\frac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p}\hat{\alpha}_i^2\right)^{\frac{1}{p}}}}} \tag{16}$$

$$\hat{K}_{HK}^{MSR} = \text{Median}\left(\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}\right) \tag{17}$$

$$\hat{K}_{HK}^{MRSR} = \text{Median}\left(\frac{1}{\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}}\right) \tag{18}$$

$$\hat{K}_{HK}^{VMSR} = \max\left(\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}\right) \tag{19}$$

$$\hat{K}_{HK}^{VMRSR} = \max\left(\frac{1}{\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}}\right) \tag{20}$$

Dorugade (2014) suggested the modification of the generalized ridge parameter in (9) by multiplying the denominator with $\lambda_{max}/2$. The estimator is defined as:

$$\hat{k}_D = \frac{2\sigma^2}{\lambda_{max}\hat{\alpha}_i^2} \tag{21}$$

where λ_{max} is the maximum eigenvalue of $X'X$.

Following Kibria (2003), Dorugade (2014) suggested the following ordinary ridge regression for the ridge parameter in (21).

$$\hat{K}_D^{HM} = \frac{2p\hat{\sigma}^2}{\lambda_{max}\sum_{i=1}^{p}\hat{\alpha}_i^2} \tag{22}$$

$$\hat{K}_D^{M} = \text{Median}\left(\frac{2\hat{\sigma}^2}{\lambda_{max}\hat{\alpha}_i^2}\right) \tag{23}$$

$$\hat{K}_D^{GM} = \frac{2\hat{\sigma}^2}{\lambda_{max}\left(\prod_{i=1}^{p}\hat{\alpha}_i^2\right)^{\frac{1}{p}}} \tag{24}$$

$$\hat{K}_{HK}^{AM} = \frac{2\hat{\sigma}^2}{\lambda_{max}p}\sum_{i=1}^{p}\frac{1}{\hat{\alpha}_i^2} \tag{25}$$

Following Dorugade (2014), Adnan et al. (2014) proposed some ridge parameters:

$$\hat{K}_{N1}^{HM} = \frac{\sqrt{5}p\hat{\sigma}^2}{\lambda_{max}\sum_{i=1}^{p}\hat{\alpha}_i^2} \tag{26}$$

$$\hat{K}_{N2}^{HM} = \frac{p\hat{\sigma}^2}{\sqrt{\lambda_{max}}\sum_{i=1}^{p}\hat{\alpha}_i^2} \tag{27}$$

$$\hat{K}_{N3}^{HM} = \frac{2p\hat{\sigma}^2}{\sum_{i=1}^{p}\left(\lambda_i^{\frac{1}{4}}\right)\sum_{i=1}^{p}\hat{\alpha}_i^2} \tag{28}$$

$$\hat{K}_{N4}^{HM} = \frac{2p\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i}\sum_{i=1}^{p}\hat{\alpha}_i^2} \tag{29}$$

The proposed ridge estimators by Dorugade (2014) and Adnan et al. (2014) are classified into different forms and various types.

Ridge Parameter Proposed by Dorugade (2014)

Dorugade (2014) proposed the ridge parameter

$$\hat{K}_{D_i} = \frac{2\hat{\sigma}^2}{\lambda_{max}\hat{\alpha}_i^2}.$$

Its estimators in the light of different forms and various types are summarized in Table 1.

Table 1. Summary of Different Forms and Various Types for $\hat{K}_{D_i} = 2\hat{\sigma}^2 / \lambda_{max}\hat{\alpha}_i^2$

Forms	Types of K			
	Original	***Reciprocal***	***Square Root***	***Reciprocal Square Root***
Fixed Maximum	$\hat{K}_D^{FMO} = \frac{2\hat{\sigma}^2}{\lambda_{max}\max\left(\hat{\alpha}_i^2\right)}$	$\hat{K}_{HK}^{FMR} = \frac{1}{\hat{K}_D^{FMO}}$	$\hat{K}_{HK}^{FMSR} = \sqrt{\hat{K}_D^{FMO}}$	$\hat{K}_D^{FMRSR} = \frac{1}{\sqrt{\hat{K}_D^{FMO}}}$
Varying Maximum	$\hat{K}_D^{VMO} = \max\left(\frac{2\hat{\sigma}^2}{\lambda_{max}\hat{\alpha}_i^2}\right)$	$\hat{K}_D^{VMR} = \max\left(\frac{1}{\hat{K}_{D_i}}\right)$	$\hat{K}_D^{VMSR} = \max\left(\sqrt{\hat{K}_{D_i}}\right)$	$\hat{K}_D^{VMRSR} = \max\left(\frac{1}{\sqrt{\hat{K}_{D_i}}}\right)$
Arithmetic Mean	$\hat{K}_D^{AMO} = \frac{2\hat{\sigma}^2}{\lambda_{max}p}\sum_{i=1}^{p}\frac{1}{\hat{\alpha}_i^2}$ *	$\hat{K}_D^{R} = \frac{1}{\hat{K}_D^{AMO}}$	$\hat{K}_D^{AMSR} = \sqrt{\hat{K}_D^{AMO}}$	$\hat{K}_D^{AMRSR} = \frac{1}{\sqrt{\hat{K}_D^{AMO}}}$
Harmonic Mean	$\hat{K}_D^{HMO} = \frac{2p\hat{\sigma}^2}{\lambda_{max}\sum_{i=1}^{p}\hat{\alpha}_i^2}$ *	$\hat{K}_D^{HMR} = \frac{1}{\hat{K}_D^{HMO}}$	$\hat{K}_D^{HMSR} = \sqrt{\hat{K}_D^{HMO}}$	$\hat{K}_D^{HMRSR} = \frac{1}{\sqrt{\hat{K}_D^{HMO}}}$
Geometric Mean	$\hat{K}_D^{GMO} = \frac{2\hat{\sigma}^2}{\lambda_{max}\left(\prod_{i=1}^{p}\hat{\alpha}_i^2\right)^{\frac{1}{p}}}$ *	$\hat{K}_D^{GMR} = \frac{1}{\hat{K}_D^{GMO}}$	$\hat{K}_D^{GMSR} = \sqrt{\hat{K}_D^{GMO}}$	$\hat{K}_D^{GMRSR} = \frac{1}{\sqrt{\hat{K}_D^{GMO}}}$
Median	$\hat{K}_D^{MO} = \text{median}\left(\frac{2\hat{\sigma}^2}{\lambda_{max}\hat{\alpha}_i^2}\right)$ *	$\hat{K}_D^{MR} = \text{median}\left(\frac{1}{\hat{K}_{D_i}}\right)$	$\hat{K}_D^{MSR} = \text{median}\left(\sqrt{\hat{K}_{D_i}}\right)$	$\hat{K}_D^{MRSR} = \text{median}\left(\frac{1}{\sqrt{\hat{K}_{D_i}}}\right)$

Notes: * Dorugade (2014); all others are proposed estimators

Ridge Parameter Proposed by Adnan et al. (2014)

Adnan et al. (2014) proposed the ridge parameter

$$\hat{K}_{AYA_i} = \frac{2\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i \hat{\alpha}_i^2}}.$$

Its estimators in the light of different forms and various types are summarized in Table 2.

Table 2. Summary of Different Forms and Various Types for $\hat{K}_{AYA_i} = 2\hat{\sigma}^2 / \sqrt{\sum_{i=1}^{p}\lambda_i \hat{\alpha}_i^2}$

Forms	Types of K			
	Original	***Reciprocal***	***Square Root***	***Reciprocal Square Root***
Fixed Maximum	$\hat{K}_{AYA}^{FMO} = \frac{2\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i \max\left(\hat{\alpha}_i^2\right)}}$	$\hat{K}_{AYA}^{FMR} = \frac{1}{\hat{K}_{AYA}^{FMO}}$	$\hat{K}_{AYA}^{FMSR} = \sqrt{\hat{K}_{AYA}^{FMO}}$	$\hat{K}_{AYA}^{FMRSR} = \frac{1}{\sqrt{\hat{K}_{AYA}^{FMO}}}$
Varying Maximum	$\hat{K}_{AYA}^{VMO} = \max\left(\frac{2\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i \hat{\alpha}_i^2}}\right)$	$\hat{K}_{AYA}^{VMR} = \max\left(\frac{1}{\hat{K}_{AYA}}\right)$	$\hat{K}_{AYA}^{VMSR} = \max\left(\sqrt{\hat{K}_{AYA}}\right)$	$\hat{K}_{AYA}^{VMRSR} = \max\left(\frac{1}{\sqrt{\hat{K}_{AYA}}}\right)$
Arithmetic Mean	$\hat{K}_{AYA}^{AMO} = \frac{2\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i}\, p \sum_{i=1}^{p}\frac{1}{\hat{\alpha}_i^2}}$	$\hat{K}_{AYA}^{AMR} = \frac{1}{\hat{K}_{AYA}^{AMO}}$	$\hat{K}_{AYA}^{AMSR} = \sqrt{\hat{K}_{AYA}^{AMO}}$	$\hat{K}_{AYA}^{AMRSR} = \frac{1}{\sqrt{\hat{K}_{AYA}^{AMO}}}$
Harmonic Mean	$\hat{K}_{AYA}^{HMO} = \frac{2p\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i}\sum_{i=1}^{p}\hat{\alpha}_i^2}$ *	$\hat{K}_{AYA}^{HMR} = \frac{1}{\hat{K}_{AYA}^{HMO}}$	$\hat{K}_{AYA}^{HMSR} = \sqrt{\hat{K}_{AYA}^{HMO}}$	$\hat{K}_{AYA}^{HMRSR} = \frac{1}{\sqrt{\hat{K}_{AYA}^{HMO}}}$
Geometric Mean	$\hat{K}_{AYA}^{GMO} = \frac{2\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i}\left(\prod_{i=1}^{p}\hat{\alpha}_i^2\right)^{\frac{1}{p}}}$	$\hat{K}_{AYA}^{GMR} = \frac{1}{\hat{K}_{AYA}^{GMO}}$	$\hat{K}_{AYA}^{GMSR} = \sqrt{\hat{K}_{AYA}^{GMO}}$	$\hat{K}_{AYA}^{GMRSR} = \frac{1}{\sqrt{\hat{K}_{AYA}^{GMO}}}$
Median	$\hat{K}_{AYA}^{MO} = \text{median}\left(\frac{2\hat{\sigma}^2}{\sqrt{\sum_{i=1}^{p}\lambda_i \hat{\alpha}_i^2}}\right)$	$\hat{K}_{AYA}^{MR} = \text{median}\left(\frac{1}{\hat{K}_{AYA}}\right)$	$\hat{K}_{AYA}^{MSR} = \text{median}\left(\sqrt{\hat{K}_{AYA}}\right)$	$\hat{K}_{AYA}^{MRSR} = \text{median}\left(\frac{1}{\sqrt{\hat{K}_{AYA}}}\right)$

Notes: * Adnan et al. (2014); all others are proposed estimators

The ridge parameter estimators in Table 1 and 2 were examined and evaluated in this study.

Monte Carlo Simulation

The considered regression model is of the form:

$$Y_t = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + B_p X_{ip} + U_t \qquad (30)$$

where $t = 1, 2, \ldots, n$; $p = 3, 7$.

The error term U_t was generated to be normally distributed with mean zero and variance σ^2, $U_t \sim N(0, \sigma^2)$. In this study, σ were taken to be 0.5, 1 and 5.

β_0 was taken to be identically zero. When $p = 3$, the values of β were chosen to be $\beta = (0.8, 0.1, 0.6)'$. When $p = 7$, the values of β were chosen to be $\beta = (0.4, 0.1, 0.6, 0.2, 0.25, 0.3, 0.53)'$. The parameter values were chosen such that $\beta'\beta = 1$ which is a common restriction in simulation studies of this type (Muniz and Kibria, 2009). We varied the sample sizes between 10, 20, 30, 40 and 50. Following McDonald and Galarneau (1975), Wichern and Churchill (1978), Gibbons (1981), Kibria (2003), Muniz and Kibria (2009), Lukman and Ayinde (2015), the explanatory variables were generated using the following equation:

$$X_{ij} = \left(1-\rho^2\right)^{\frac{1}{2}} Z_{ij} + \rho Z_{ip},\ i = 1,2,3,\ldots,n,\ \ j = 1,2,\ldots,p. \tag{31}$$

where Z_{ij} is independent standard normal distribution with mean zero and unit variance, ρ is the correlation between any two explanatory variables and p is the number of explanatory variables. The number of explanatory variable (p) is taken to be three (3) and seven (7). The value of ρ is taken as 0.95, 0.99 respectively. Three different values of σ, 0.5, 1 and 5, were also used. The experiment is replicated 1,000 times. The ridge parameter estimators are evaluated using mean square error (MSE).

Results

The results of the simulation are presented in Table 3 and 4. These tables provide the results of the estimated mean square error of the ridge parameter when the number of regressors is three (3) and seven (7) respectively. The mean square error increases as the multicollinearity level increases. Across each multicollinearity level, the mean square error decreases as the sample sizes increase from 10 to 50, while increasing the number of regressors increases the estimated MSE. However, it is observed that the ridge estimators based on K_{AYA} performed consistently better than K_D. Occasionally, this method performs better than K_{AYA}. For instance, estimators $\hat{K}_D^{VMSR}$ and $\hat{K}_D^{AMSR}$ perform consistently well over estimators based on K_{AYA} especially when the number of regressors increases to seven (7), and when the number of regressors is three (3), especially when $n \leq 20$. This can be seen in Figure 1 and 2. The following ridge parameter

estimators based on K_{AYA}: $\hat{K}_{AYA}^{FMSR}$, $\hat{K}_{AYA}^{HMO}$, $\hat{K}_{AYA}^{FMO}$, $\hat{K}_{AYA}^{HMSR}$, $\hat{K}_{AYA}^{GMO}$, $\hat{K}_{AYA}^{MO}$, and $\hat{K}_{AYA}^{GMSR}$ performed best when compared to others. All but $\hat{K}_{AYA}^{HMO}$ are proposed in this study. When $p = 3$, $\hat{K}_{AYA}^{FMSR}$ performs better than the existing ridge parameter $\hat{K}_{AYA}^{HMO}$, while $\hat{K}_{AYA}^{HMO}$ performs better than $\hat{K}_{AYA}^{FMSR}$ when $p = 7$. The estimators considered best in this study have the least MSE when compared to others. The proposed estimators perform better than the existing estimators based on K_D.

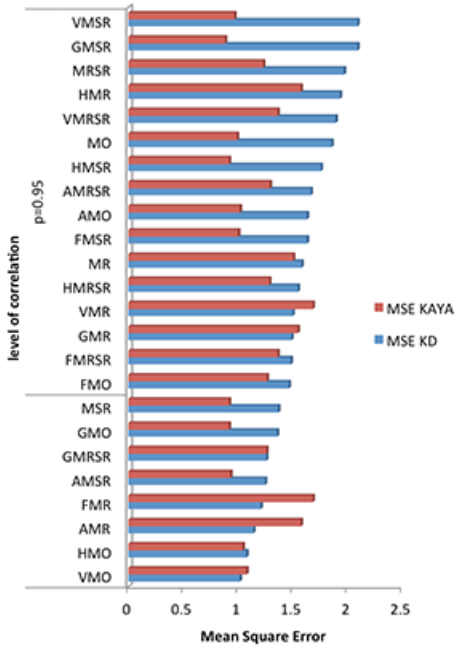


Figure 1. Graphical Illustration when $n = 20$, $\sigma^2 = 0.25$, $p = 3$

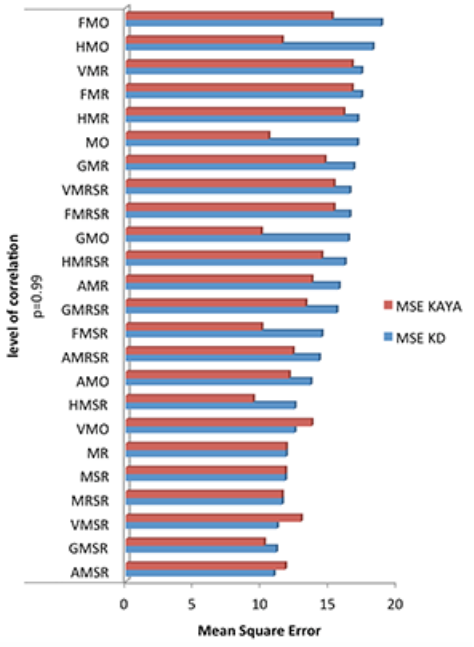


Figure 2. Graphical Illustration when $n = 50$, $\sigma^2 = 0.25$, $p = 7$

Table 3. Estimated Mean Square Error of ridge parameter when $p = 3$

	$p = 3, \sigma = 0.5, \rho = 0.95$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	3.501	2.297	1.757	1.265	1.633	1.224	0.877	0.750	0.680	0.611
FMR	4.470	4.049	2.091	1.683	2.149	1.722	1.191	0.795	1.017	0.631
FMSR	2.340	1.757	1.369	1.004	1.309	0.941	0.789	0.627	0.635	0.532
FMRSR	3.904	3.550	1.630	1.363	1.690	1.390	0.805	0.595	0.649	0.456
VMO	2.374	2.263	1.248	1.078	1.235	1.039	0.705	0.572	0.581	0.471
VMR	4.470	4.049	2.091	1.683	2.149	1.722	1.191	0.795	1.017	0.631
VMSR	2.004	2.289	1.019	0.969	1.000	0.948	0.607	0.493	0.514	0.407
VMRSR	3.904	3.550	1.630	1.363	1.690	1.390	0.805	0.595	0.649	0.456
AMO	2.589	2.033	1.358	1.019	1.337	0.987	0.755	0.571	0.611	0.477
AMR	4.092	3.659	1.856	1.575	1.889	1.532	1.016	0.760	0.843	0.592
AMSR	1.995	2.088	1.079	0.932	1.054	0.906	0.652	0.506	0.546	0.426
AMRSR	3.532	3.107	1.465	1.293	1.490	1.242	0.725	0.630	0.572	0.500
HMO	3.184	1.863	1.665	1.045	1.579	1.027	0.865	0.666	0.674	0.559
HMR	4.322	3.889	1.968	1.573	2.037	1.606	1.102	0.724	0.938	0.568
HMSR	2.098	1.723	1.259	0.920	1.215	0.868	0.759	0.575	0.618	0.494
HMRSR	3.789	3.380	1.548	1.287	1.614	1.302	0.754	0.574	0.602	0.448
GMO	2.782	1.762	1.483	0.919	1.447	0.903	0.821	0.562	0.651	0.476
GMR	4.206	3.747	1.892	1.546	1.956	1.535	1.043	0.730	0.882	0.571
GMSR	1.963	1.855	1.142	0.883	1.109	0.845	0.709	0.521	0.588	0.447
GMRSR	3.665	3.215	1.489	1.260	1.546	1.242	0.727	0.592	0.577	0.468
MO	2.930	1.856	1.581	0.992	1.508	0.965	0.837	0.618	0.658	0.508
MR	4.243	3.710	1.931	1.505	2.007	1.532	1.105	0.732	0.938	0.586
MSR	2.045	1.910	1.209	0.920	1.162	0.874	0.733	0.549	0.600	0.467
MRSR	3.675	3.192	1.499	1.230	1.567	1.230	0.754	0.580	0.605	0.461

Table 3, continued.

	$p = 3, \sigma = 0.5, \rho = 0.99$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	18.092	11.590	8.556	5.451	8.606	5.586	4.642	3.313	3.604	2.739
FMR	23.328	22.984	9.951	9.565	10.616	10.215	5.488	5.093	4.354	3.988
FMSR	9.438	9.998	4.895	4.092	5.062	4.119	3.310	2.287	2.815	1.905
FMRSR	22.714	22.414	9.400	9.056	10.082	9.728	5.018	4.634	3.932	3.534
VMO	11.597	16.282	5.494	6.071	6.106	6.470	3.466	2.787	2.872	2.165
VMR	23.328	22.984	9.951	9.565	10.616	10.215	5.488	5.093	4.354	3.988
VMSR	14.796	18.237	5.274	6.626	5.578	7.151	2.623	2.900	2.174	2.099
VMRSR	22.714	22.414	9.400	9.056	10.082	9.728	5.018	4.634	3.932	3.534
AMO	11.942	13.339	6.064	5.189	6.616	5.595	3.748	2.659	3.087	2.144
AMR	22.669	21.913	9.565	8.967	10.164	9.469	5.173	4.703	4.038	3.562
AMSR	13.197	16.753	4.928	5.998	5.217	6.536	2.692	2.669	2.292	1.972
AMRSR	22.012	21.079	8.921	8.200	9.590	8.753	4.695	4.075	3.603	2.997
HMO	15.724	9.207	7.667	4.343	7.982	4.490	4.469	2.739	3.517	2.309
HMR	23.199	22.808	9.808	9.399	10.482	10.073	5.361	4.965	4.240	3.868
HMSR	9.301	11.554	4.433	4.382	4.564	4.508	2.997	2.214	2.595	1.760
HMRSR	22.602	22.215	9.291	8.846	9.992	9.529	4.936	4.450	3.855	3.358
GMO	12.880	9.977	6.567	4.300	6.973	4.520	4.091	2.440	3.318	2.024
GMR	23.011	22.525	9.658	9.191	10.324	9.846	5.248	4.809	4.153	3.710
GMSR	11.006	14.465	4.495	5.213	4.643	5.625	2.745	2.381	2.387	1.792
GMRSR	22.407	21.806	9.128	8.529	9.840	9.186	4.823	4.231	3.753	3.156
MO	13.018	11.525	6.768	4.807	7.190	5.032	4.263	2.569	3.396	2.121
MR	22.980	22.395	9.662	9.091	10.321	9.762	5.264	4.760	4.171	3.693
MSR	11.634	14.736	4.766	5.443	4.932	5.955	2.864	2.499	2.464	1.856
MRSR	22.373	21.717	9.106	8.456	9.822	9.103	4.822	4.177	3.758	3.124

Table 3, continued.

	$p = 3, \sigma = 1, \rho = 0.95$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	3.730	2.432	1.811	1.306	1.675	1.249	0.894	0.763	0.690	0.620
FMR	4.650	4.224	2.228	1.806	2.151	1.726	1.190	0.794	1.015	0.630
FMSR	2.488	1.864	1.413	1.039	1.340	0.960	0.804	0.638	0.644	0.539
FMRSR	4.077	3.716	1.747	1.457	1.695	1.396	0.805	0.597	0.649	0.459
VMO	2.516	2.394	1.291	1.127	1.259	1.052	0.720	0.582	0.584	0.472
VMR	4.650	4.224	2.228	1.806	2.151	1.726	1.190	0.794	1.015	0.630
VMSR	2.127	2.426	1.063	1.021	1.020	0.961	0.617	0.499	0.520	0.410
VMRSR	4.077	3.716	1.747	1.457	1.695	1.396	0.805	0.597	0.649	0.459
AMO	2.749	2.150	1.407	1.060	1.367	1.001	0.771	0.582	0.616	0.478
AMR	4.309	3.860	1.975	1.663	1.900	1.548	1.020	0.768	0.843	0.598
AMSR	2.117	2.216	1.123	0.977	1.078	0.920	0.663	0.513	0.553	0.430
AMRSR	3.725	3.288	1.558	1.361	1.496	1.257	0.730	0.639	0.574	0.506
HMO	3.386	1.969	1.716	1.080	1.619	1.045	0.882	0.676	0.685	0.566
HMR	4.497	4.059	2.102	1.686	2.040	1.611	1.101	0.725	0.937	0.569
HMSR	2.229	1.829	1.299	0.954	1.243	0.885	0.773	0.584	0.627	0.501
HMRSR	3.956	3.539	1.658	1.372	1.618	1.309	0.754	0.579	0.602	0.452
GMO	2.945	1.868	1.535	0.961	1.483	0.922	0.835	0.570	0.662	0.481
GMR	4.395	3.936	2.018	1.642	1.956	1.541	1.044	0.735	0.882	0.574
GMSR	2.081	1.970	1.184	0.923	1.135	0.862	0.721	0.528	0.597	0.453
GMRSR	3.837	3.381	1.589	1.335	1.549	1.250	0.729	0.599	0.578	0.473
MO	3.110	1.966	1.634	1.028	1.543	0.982	0.853	0.628	0.668	0.515
MR	4.411	3.853	2.059	1.602	2.010	1.537	1.104	0.734	0.937	0.587
MSR	2.173	2.036	1.250	0.958	1.188	0.890	0.746	0.557	0.609	0.473
MRSR	3.834	3.333	1.601	1.305	1.571	1.238	0.755	0.586	0.606	0.465

Table 3, continued.

	$p = 3, \sigma = 1, \rho = 0.99$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	19.320	12.313	8.834	5.679	8.819	5.696	4.733	3.366	3.659	2.771
FMR	24.239	23.888	10.731	10.337	10.647	10.247	5.495	5.099	4.357	3.991
FMSR	10.025	10.572	5.094	4.299	5.168	4.187	3.368	2.318	2.854	1.926
FMRSR	23.617	23.307	10.161	9.794	10.115	9.758	5.025	4.638	3.936	3.537
VMO	12.249	17.173	5.723	6.502	6.221	6.549	3.509	2.806	2.902	2.180
VMR	24.239	23.888	10.731	10.337	10.647	10.247	5.495	5.099	4.357	3.991
VMSR	15.584	19.109	5.605	7.122	5.650	7.209	2.648	2.914	2.191	2.112
VMRSR	23.617	23.307	10.161	9.794	10.115	9.758	5.025	4.638	3.936	3.537
AMO	12.630	14.109	6.275	5.487	6.741	5.673	3.806	2.676	3.120	2.161
AMR	23.633	22.899	10.276	9.604	10.198	9.528	5.186	4.710	4.041	3.567
AMSR	13.939	17.597	5.190	6.413	5.290	6.596	2.723	2.684	2.312	1.985
AMRSR	22.911	21.970	9.631	8.816	9.639	8.791	4.702	4.083	3.607	3.001
HMO	16.754	9.756	7.941	4.546	8.168	4.569	4.553	2.776	3.569	2.333
HMR	24.107	23.707	10.588	10.161	10.513	10.104	5.367	4.970	4.243	3.871
HMSR	9.865	12.191	4.630	4.626	4.652	4.572	3.047	2.241	2.629	1.778
HMRSR	23.501	23.099	10.045	9.567	10.024	9.558	4.942	4.453	3.859	3.361
GMO	13.612	10.593	6.819	4.498	7.099	4.591	4.167	2.466	3.365	2.040
GMR	23.903	23.382	10.441	9.922	10.361	9.899	5.259	4.815	4.160	3.715
GMSR	11.654	15.236	4.708	5.533	4.714	5.684	2.786	2.401	2.416	1.806
GMRSR	23.293	22.663	9.872	9.222	9.875	9.221	4.832	4.237	3.758	3.161
MO	13.811	12.216	7.015	5.072	7.339	5.106	4.338	2.598	3.442	2.143
MR	23.880	23.264	10.423	9.789	10.353	9.795	5.270	4.763	4.176	3.698
MSR	12.330	15.539	5.007	5.781	5.009	6.012	2.905	2.521	2.494	1.873
MRSR	23.259	22.573	9.839	9.130	9.853	9.130	4.827	4.180	3.762	3.129

Table 3, continued.

	$p = 3, \sigma = 5, \rho = 0.95$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	11.368	7.438	3.629	2.802	2.943	1.823	1.443	1.109	1.019	0.857
FMR	10.621	10.033	6.060	5.043	2.220	1.935	1.172	0.872	0.997	0.690
FMSR	7.609	5.767	2.910	2.281	2.232	1.451	1.265	0.940	0.937	0.750
FMRSR	9.739	9.054	4.878	3.849	1.820	1.663	0.835	0.762	0.672	0.589
VMO	7.360	8.088	2.773	2.920	1.880	1.335	1.014	0.712	0.805	0.580
VMR	10.621	10.033	6.060	5.043	2.220	1.935	1.172	0.872	0.997	0.690
VMSR	6.811	7.703	2.385	2.638	1.505	1.251	0.884	0.641	0.719	0.528
VMRSR	9.739	9.054	4.878	3.849	1.820	1.663	0.835	0.762	0.672	0.589
AMO	7.511	7.398	3.017	2.465	2.125	1.309	1.123	0.732	0.870	0.604
AMR	10.043	9.853	5.115	3.460	2.193	2.206	1.145	1.113	0.912	0.821
AMSR	6.579	7.210	2.448	2.377	1.644	1.242	0.974	0.680	0.777	0.565
AMRSR	8.675	8.335	3.880	2.948	1.791	1.824	0.889	0.973	0.683	0.731
HMO	9.971	5.816	3.471	2.400	2.752	1.446	1.406	0.928	1.006	0.752
HMR	10.380	9.682	5.817	4.507	2.124	1.926	1.091	0.891	0.927	0.689
HMSR	6.650	5.618	2.704	2.158	2.021	1.297	1.201	0.840	0.906	0.685
HMRSR	9.452	8.641	4.562	3.485	1.760	1.643	0.803	0.797	0.640	0.615
GMO	7.658	6.150	3.220	2.209	2.420	1.247	1.293	0.749	0.956	0.619
GMR	9.931	9.221	5.533	3.973	2.107	2.036	1.080	1.007	0.892	0.757
GMSR	6.115	6.311	2.529	2.169	1.803	1.220	1.099	0.736	0.852	0.609
GMRSR	8.890	8.085	4.239	3.182	1.739	1.690	0.824	0.877	0.644	0.669
MO	8.041	6.399	3.326	2.325	2.588	1.354	1.338	0.844	0.975	0.683
MR	9.931	8.925	5.641	4.182	2.105	1.882	1.116	0.942	0.941	0.721
MSR	6.330	6.458	2.613	2.193	1.913	1.285	1.146	0.788	0.877	0.646
MRSR	8.852	7.938	4.375	3.302	1.716	1.602	0.824	0.833	0.655	0.639

Table 3, continued.

	$p = 3, \sigma = 5, \rho = 0.99$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	59.692	37.770	18.672	13.998	15.022	8.319	7.487	4.664	5.332	3.620
FMR	53.657	53.218	32.462	31.863	11.276	10.893	5.636	5.244	4.434	4.088
FMSR	30.381	30.342	12.064	11.318	7.969	5.753	4.993	3.111	3.975	2.487
FMRSR	52.809	52.248	31.367	30.231	10.761	10.376	5.175	4.763	4.032	3.644
VMO	38.646	46.837	14.345	22.905	8.345	7.902	4.911	3.312	3.799	2.541
VMR	53.657	53.218	32.462	31.863	11.276	10.893	5.636	5.244	4.434	4.088
VMSR	42.916	47.776	17.149	23.077	7.080	8.247	3.447	3.340	2.766	2.443
VMRSR	52.809	52.248	31.367	30.231	10.761	10.376	5.175	4.763	4.032	3.644
AMO	37.226	42.827	13.573	17.671	9.686	7.071	5.585	3.234	4.197	2.570
AMR	53.188	53.091	31.009	26.815	10.952	10.668	5.464	5.240	4.288	4.035
AMSR	40.230	45.767	14.585	20.156	6.903	7.713	3.689	3.163	2.993	2.351
AMRSR	49.835	48.112	28.792	25.131	10.257	9.549	4.890	4.452	3.789	3.374
HMO	50.356	29.122	17.437	11.743	13.310	6.311	7.033	3.630	5.128	2.922
HMR	53.471	52.979	32.281	31.474	11.141	10.762	5.512	5.110	4.328	3.984
HMSR	29.248	33.662	11.305	12.463	6.897	5.957	4.396	2.882	3.602	2.239
HMRSR	52.585	51.797	31.043	29.510	10.654	10.172	5.076	4.586	3.949	3.490
GMO	35.454	34.552	15.197	11.493	11.040	6.049	6.262	3.140	4.669	2.489
GMR	52.980	51.557	31.959	30.408	10.994	10.583	5.417	5.054	4.261	3.947
GMSR	34.408	41.170	11.755	15.542	6.585	6.913	3.917	2.952	3.220	2.206
GMRSR	51.759	50.074	30.358	28.030	10.483	9.835	4.959	4.430	3.852	3.358
MO	37.649	36.864	15.980	12.632	11.137	6.681	6.483	3.275	4.891	2.644
MR	53.025	51.491	32.038	30.644	10.985	10.394	5.437	4.922	4.260	3.795
MSR	35.513	41.459	11.932	14.442	6.907	7.242	4.045	3.064	3.376	2.305
MRSR	51.761	50.035	30.626	28.647	10.446	9.710	4.952	4.354	3.838	3.263

Table 4. Estimated Mean Square Error of ridge parameter when $p = 7$

	$p = 7, \sigma = 0.5, \rho = 0.95$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	76.120	47.232	7.028	5.439	4.881	4.079	2.493	2.267	2.262	2.018
FMR	81.940	81.594	6.749	6.071	5.540	4.870	2.855	2.186	2.544	1.960
FMSR	36.541	45.788	5.381	3.873	4.109	3.051	2.316	1.900	2.072	1.655
FMRSR	81.165	80.643	5.940	5.289	4.789	4.137	2.192	1.660	1.960	1.584
VMO	66.682	76.133	4.525	5.406	3.600	4.332	1.837	2.072	1.625	1.918
VMR	81.940	81.594	6.749	6.071	5.540	4.870	2.855	2.186	2.544	1.960
VMSR	72.487	76.644	4.136	4.789	3.209	3.754	1.597	1.636	1.439	1.530
VMRSR	81.165	80.643	5.940	5.289	4.789	4.137	2.192	1.660	1.960	1.584
AMO	58.101	71.296	4.656	4.586	3.658	3.671	1.996	1.733	1.750	1.601
AMR	78.493	74.787	5.534	5.222	4.306	3.761	2.071	1.971	1.842	1.835
AMSR	68.811	74.331	3.983	4.278	3.119	3.342	1.736	1.504	1.528	1.405
AMRSR	78.005	74.811	4.786	4.485	3.667	3.320	1.661	1.773	1.535	1.637
HMO	57.198	33.682	6.594	4.036	4.731	3.217	2.463	1.946	2.226	1.666
HMR	81.751	81.162	6.439	5.582	5.263	4.403	2.644	1.833	2.335	1.712
HMSR	43.538	57.236	4.637	3.498	3.687	2.736	2.199	1.669	1.939	1.445
HMRSR	80.853	80.051	5.639	4.879	4.515	3.768	1.978	1.534	1.790	1.492
GMO	42.525	58.346	5.592	3.572	4.256	2.806	2.355	1.537	2.079	1.331
GMR	81.271	79.981	6.073	4.937	4.957	3.766	2.405	1.683	2.099	1.588
GMSR	60.770	69.328	4.072	3.645	3.258	2.818	2.023	1.475	1.743	1.321
GMRSR	80.170	78.605	5.239	4.455	4.146	3.378	1.767	1.534	1.616	1.456
MO	45.003	60.574	5.992	3.791	4.438	2.983	2.397	1.702	2.122	1.437
MR	77.079	77.064	4.594	4.616	3.422	3.432	1.983	1.987	1.858	1.862
MSR	74.023	73.944	4.031	4.027	3.163	3.162	1.347	1.347	1.310	1.310
MRSR	76.542	76.518	4.168	4.172	3.130	3.132	1.741	1.744	1.611	1.613

Table 4, continued.

	$p = 7, \sigma = 0.5, \rho = 0.99$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	447.805	273.075	37.946	28.276	27.117	21.641	13.239	11.463	12.476	10.244
FMR	485.223	485.064	34.950	34.397	29.708	29.129	14.293	13.683	13.301	12.705
FMSR	258.034	346.812	22.494	18.038	18.429	14.253	10.843	7.742	9.695	6.900
FMRSR	484.613	484.223	34.080	33.385	28.880	28.145	13.521	12.635	12.567	11.742
VMO	419.503	465.165	23.743	28.568	20.208	24.458	9.430	10.457	9.084	10.228
VMR	485.223	485.064	34.950	34.397	29.708	29.129	14.293	13.683	13.301	12.705
VMSR	459.749	473.060	24.503	28.718	20.411	24.204	8.218	9.863	8.292	9.733
VMRSR	484.613	484.223	34.080	33.385	28.880	28.145	13.521	12.635	12.567	11.742
AMO	368.598	442.992	24.438	24.674	20.076	21.319	10.316	8.959	9.659	8.950
AMR	482.205	476.562	32.652	30.640	27.796	24.643	12.890	10.899	11.882	10.120
AMSR	447.205	465.894	22.355	26.311	18.496	22.178	8.094	8.673	7.963	8.840
AMRSR	481.236	477.172	31.482	28.781	26.510	23.579	11.709	9.718	10.819	9.212
HMO	327.951	196.248	34.716	20.103	25.809	16.181	12.964	9.262	12.106	7.938
HMR	485.147	484.810	34.688	33.978	29.441	28.742	14.020	13.272	13.029	12.282
HMSR	331.301	401.383	19.395	19.438	15.862	15.458	9.612	6.985	8.450	6.648
HMRSR	484.361	483.812	33.801	32.646	28.625	27.393	13.264	11.887	12.313	11.078
GMO	258.937	377.570	28.176	19.582	22.283	16.053	12.173	7.586	11.000	7.152
GMR	484.816	484.036	34.340	32.952	29.120	27.714	13.791	12.336	12.782	11.289
GMSR	418.634	449.401	19.796	22.825	15.970	18.870	8.533	7.266	7.694	7.517
GMRSR	483.824	482.511	33.300	31.360	28.138	26.069	12.861	10.917	11.877	10.176
MO	286.980	388.382	29.404	20.143	23.176	16.568	12.488	8.068	11.506	7.514
MR	481.859	481.850	29.872	29.865	23.813	23.805	9.406	9.406	8.879	8.879
MSR	464.829	464.318	26.449	26.407	22.389	22.372	8.670	8.663	8.806	8.802
MRSR	480.163	480.143	29.045	29.028	23.410	23.399	9.042	9.038	8.746	8.743

Table 4, continued.

	$p = 7, \sigma = 1, \rho = 0.95$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	81.355	51.963	7.225	5.692	5.002	4.184	2.546	2.311	2.298	2.048
FMR	101.096	100.762	7.808	7.122	5.690	5.012	2.866	2.192	2.559	1.972
FMSR	41.344	54.271	5.599	4.132	4.213	3.129	2.364	1.935	2.103	1.678
FMRSR	100.314	99.768	6.965	6.277	4.927	4.254	2.200	1.670	1.973	1.598
VMO	82.129	94.427	5.134	6.301	3.688	4.469	1.879	2.077	1.656	1.959
VMR	101.096	100.762	7.808	7.122	5.690	5.012	2.866	2.192	2.559	1.972
VMSR	90.071	95.087	4.777	5.626	3.295	3.870	1.631	1.651	1.456	1.559
VMRSR	100.314	99.768	6.965	6.277	4.927	4.254	2.200	1.670	1.973	1.598
AMO	70.252	88.388	5.085	5.329	3.746	3.782	2.048	1.756	1.766	1.639
AMR	97.638	93.084	6.270	5.636	4.394	3.901	2.088	2.008	1.877	1.876
AMSR	85.460	92.346	4.485	5.009	3.208	3.439	1.775	1.528	1.538	1.428
AMRSR	96.866	93.247	5.506	4.960	3.742	3.418	1.678	1.805	1.570	1.676
HMO	62.106	37.820	6.801	4.291	4.849	3.298	2.515	1.980	2.260	1.688
HMR	100.913	100.326	7.485	6.594	5.411	4.528	2.655	1.841	2.350	1.727
HMSR	51.275	69.899	4.858	3.835	3.779	2.802	2.244	1.699	1.968	1.464
HMRSR	99.984	99.135	6.639	5.797	4.644	3.870	1.985	1.549	1.803	1.508
GMO	49.184	72.188	5.809	4.021	4.356	2.872	2.402	1.560	2.106	1.341
GMR	100.412	99.036	7.093	5.768	5.099	3.863	2.417	1.704	2.110	1.612
GMSR	75.548	86.521	4.371	4.189	3.333	2.884	2.062	1.498	1.765	1.334
GMRSR	99.227	97.505	6.178	5.224	4.262	3.462	1.778	1.557	1.629	1.477
MO	52.630	75.436	6.197	4.179	4.547	3.056	2.447	1.731	2.155	1.459
MR	95.434	95.380	5.109	5.127	3.494	3.504	2.026	2.031	1.887	1.892
MSR	92.275	92.184	4.751	4.746	3.246	3.244	1.363	1.363	1.326	1.326
MRSR	95.045	95.014	4.754	4.754	3.200	3.202	1.777	1.779	1.635	1.638

Table 4, continued.

	$p = 7, \sigma = 1, \rho = 0.99$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	479.139	301.137	39.063	29.750	27.804	22.235	13.521	11.688	12.671	10.387
FMR	601.450	601.301	40.901	40.362	30.487	29.903	14.363	13.750	13.399	12.801
FMSR	303.518	422.962	23.865	20.055	18.910	14.607	11.059	7.873	9.835	6.988
FMRSR	600.849	600.446	40.020	39.319	29.651	28.898	13.588	12.689	12.663	11.829
VMO	521.305	578.507	27.350	33.856	20.911	25.040	9.655	10.539	9.117	10.329
VMR	601.450	601.301	40.901	40.362	30.487	29.903	14.363	13.750	13.399	12.801
VMSR	572.441	587.779	29.099	34.101	20.959	24.853	8.356	9.942	8.340	9.813
VMRSR	600.849	600.446	40.020	39.319	29.651	28.898	13.588	12.689	12.663	11.829
AMO	456.594	552.428	26.595	29.195	20.724	21.906	10.561	9.080	9.685	9.035
AMR	599.015	594.110	39.025	36.016	28.368	25.098	12.893	10.930	11.870	10.240
AMSR	557.880	579.699	26.220	31.327	19.004	22.800	8.244	8.765	8.007	8.906
AMRSR	597.876	593.860	37.571	34.529	27.070	24.034	11.731	9.725	10.879	9.258
HMO	357.282	220.799	35.927	21.611	26.467	16.608	13.235	9.418	12.293	8.041
HMR	601.379	601.052	40.645	39.929	30.218	29.511	14.089	13.332	13.127	12.373
HMSR	402.765	496.974	21.078	22.475	16.253	15.805	9.793	7.083	8.569	6.727
HMRSR	600.587	600.009	39.730	38.547	29.390	28.126	13.327	11.932	12.407	11.160
GMO	307.081	470.214	29.384	22.461	22.805	16.390	12.417	7.694	11.182	7.217
GMR	601.046	600.232	40.284	38.886	29.891	28.429	13.860	12.374	12.885	11.378
GMSR	523.625	560.779	22.637	27.212	16.298	19.295	8.679	7.345	7.796	7.590
GMRSR	600.000	598.580	39.196	37.154	28.882	26.745	12.917	10.954	11.972	10.257
MO	344.556	486.174	30.481	23.214	23.721	16.948	12.738	8.169	11.673	7.610
MR	597.746	597.734	35.219	35.201	24.426	24.419	9.471	9.472	8.965	8.966
MSR	578.893	578.320	31.713	31.665	22.965	22.948	8.725	8.718	8.889	8.885
MRSR	595.905	595.883	34.454	34.434	24.003	23.992	9.094	9.090	8.820	8.817

Table 4, continued.

	$p = 7, \sigma = 5, \rho = 0.95$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	245.847	206.324	13.648	12.477	8.925	7.687	4.247	3.636	3.406	2.937
FMR	702.157	701.896	38.766	37.977	10.527	9.586	3.195	2.433	3.243	2.535
FMSR	200.188	334.411	11.878	10.816	7.614	5.710	3.873	3.014	3.086	2.405
FMRSR	701.135	700.018	37.386	35.751	9.385	8.006	2.435	2.084	2.536	2.116
VMO	610.902	681.797	23.315	33.177	6.684	8.515	2.612	2.460	2.307	2.557
VMR	702.157	701.896	38.766	37.977	10.527	9.586	3.195	2.433	3.243	2.535
VMSR	655.825	681.263	24.043	30.872	5.936	7.314	2.321	2.025	2.045	2.081
VMRSR	701.135	700.018	37.386	35.751	9.385	8.006	2.435	2.084	2.536	2.116
AMO	507.349	652.570	17.667	28.221	6.640	7.108	3.127	2.187	2.559	2.192
AMR	671.962	627.505	30.695	20.513	7.973	6.736	2.732	3.339	2.492	2.740
AMSR	627.882	668.286	19.588	27.354	5.634	6.357	2.707	2.008	2.217	1.947
AMRSR	678.095	654.138	28.502	21.283	6.708	5.920	2.343	2.966	2.130	2.427
HMO	219.097	173.903	13.235	10.792	8.651	5.956	4.166	2.932	3.339	2.376
HMR	702.008	701.327	38.330	36.735	10.155	8.469	2.951	2.289	3.006	2.263
HMSR	307.207	481.565	11.101	12.850	6.747	4.953	3.636	2.568	2.877	2.085
HMRSR	700.344	698.117	36.401	33.452	8.766	7.010	2.225	2.161	2.320	2.041
GMO	272.739	544.532	11.967	18.062	7.653	5.074	3.925	2.192	3.104	1.884
GMR	700.807	695.054	37.026	30.008	9.481	6.656	2.678	2.595	2.697	2.241
GMSR	555.757	634.786	13.407	21.094	5.773	5.074	3.309	2.187	2.580	1.882
GMRSR	696.388	688.388	33.710	27.725	7.809	5.970	2.106	2.400	2.122	2.072
MO	314.950	575.882	12.144	18.589	7.980	5.216	4.041	2.530	3.195	2.061
MR	672.820	672.433	17.880	17.781	6.061	6.084	3.521	3.530	2.814	2.821
MSR	668.530	668.190	27.900	27.872	6.048	6.045	1.790	1.791	1.802	1.802
MRSR	674.310	674.097	20.956	20.872	5.432	5.436	2.985	2.991	2.395	2.399

Table 4, continued.

	$p = 7, \sigma = 5, \rho = 0.99$									
	$n = 10$		$n = 20$		$n = 30$		$n = 40$		$n = 50$	
Methods	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$	$\hat{K}_D$	$\hat{K}_{AYA}$
FMO	1465.681	1221.250	74.342	67.324	49.754	41.369	22.547	18.404	18.808	15.170
FMR	4242.117	4242.026	218.709	218.292	55.081	54.422	16.355	15.626	17.341	16.648
FMSR	1810.703	2897.433	59.291	71.107	34.486	25.768	17.728	11.619	14.416	9.990
FMRSR	4241.490	4240.837	217.652	216.573	54.035	52.742	15.463	14.173	16.487	15.313
VMO	3993.568	4193.263	140.741	198.687	36.040	46.002	13.257	12.360	12.430	13.682
VMR	4242.117	4242.026	218.709	218.292	55.081	54.422	16.355	15.626	17.341	16.648
VMSR	4149.675	4204.280	173.098	199.007	36.755	45.257	10.669	11.592	11.141	12.899
VMRSR	4241.490	4240.837	217.652	216.573	54.035	52.742	15.463	14.173	16.487	15.313
AMO	3570.762	4095.430	100.596	173.300	35.184	39.306	15.756	11.148	13.609	12.012
AMR	4212.933	4143.105	210.922	190.956	50.930	43.785	15.140	14.313	15.691	13.685
AMSR	4079.126	4172.108	150.398	186.934	32.406	40.952	11.345	10.514	10.893	11.728
AMRSR	4225.677	4199.721	208.595	194.082	48.556	41.904	13.481	12.148	14.228	12.291
HMO	1297.400	1030.908	71.579	57.889	47.574	30.452	21.850	13.783	18.180	11.491
HMR	4242.073	4241.811	218.503	217.683	54.787	53.806	16.070	14.931	17.057	16.032
HMSR	2733.910	3548.233	64.883	106.353	28.803	26.953	15.284	9.776	12.436	9.368
HMRSR	4241.023	4239.830	217.053	214.838	53.558	51.219	15.076	13.201	16.126	14.403
GMO	1907.678	3525.788	64.386	116.466	40.055	28.037	19.950	10.384	16.380	9.948
GMR	4241.579	4239.498	217.808	214.219	54.328	51.297	15.782	13.679	16.767	14.639
GMSR	3848.564	4067.859	112.552	163.774	27.263	33.591	13.008	9.372	11.073	10.181
GMRSR	4238.586	4232.948	214.991	208.642	52.448	48.033	14.490	12.179	15.517	13.231
MO	2077.145	3612.882	66.849	123.869	40.641	28.986	20.697	11.191	17.048	10.500
MR	4230.358	4230.299	197.452	197.302	41.674	41.646	11.876	11.896	11.794	11.799
MSR	4154.180	4152.100	190.071	189.885	41.694	41.662	10.327	10.321	11.743	11.739
MRSR	4222.197	4222.106	196.988	196.881	41.815	41.785	10.818	10.821	11.499	11.496

Conclusion

In this study, ridge parameters proposed by Dorugade (2014) and Adnan et al. (2014) are classified into different forms and various types following the idea of Lukman and Ayinde (2015), and some new ridge parameters are proposed. The performances of these estimators are evaluated through Monte Carlo Simulation, where levels of multicollinearity, sample sizes, number of regressors and error variances have been varied. The performance evaluation was done using the mean square error. The proposed estimators generally have the least minimum square error when compared to others.

References

Adnan, K., Yasin, A. & Asir, G. (2014). Some new modifications of Kibria's and Dorugade's methods: An application to Turkish GDP data. *Journal*

of the Association of Arab Universities for Basic and Applied Sciences. *20*, 89-99. doi: 10.1016/j.jaubas.2014.08.005

Alkhamisi, M., Khalaf, G. & Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics- Theory and Methods*, *35*(11), 2005-2020. doi: 10.1080/03610920600762905

Alkhamisi, M. & Shukur, G. (2007). A Monte Carlo study of recent ridge parameters. *Communications in Statistics- Simulation and Computation, 36*(3), 535-547. doi: 10.1080/03610910701208619

Dorugade, A. V. (2014). New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, *15*, 94-99. doi: 10.1016/j.jaubas.2013.03.005

Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, *76*(373), 131-139. doi: 10.1080/01621459.1981.10477619

Hocking, R., Speed, F. M. & Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics*, *18*(4), 425-437. doi: 10.1080/00401706.1976.10489474

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, *12*(1), 55-67. doi: 10.1080/00401706.1970.10488634

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: Some simulation. *Communications in Statistics – Simulation and Computation*, *4*(2), 105–123. doi: 10.1080/03610917508548342

Khalaf, G. & Shukur, G. (2005). Choosing ridge parameters for regression problems. *Communications in Statistics- Theory and Methods*, *34*(5), 1177-1182. doi: 10.1081/sta-200056836

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation*, *32*(2), 419-435. doi: 10.1081/sac-120017499

Koutsoyiannis, A. (2003). *Theory of Econometrics* (2nd Ed). Basingstoke, UK: Palgrave.

Lawless, J. F. & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods*, 5(4), 307-323. doi: 10.1080/03610927608827353

Lukman, A. F. & Ayinde, K. (2015). Review and classification of the Ridge Parameter Estimation Techniques. *Hacettepe Journal of Mathematics and Statistics, 46*(113), 1-1. doi: 10.15672/hjms.201815671

Mansson, K., Shukur, G. & Kibria, B. M. G. (2010). A simulation study of some ridge regression estimators under different distributional assumptions. *Communications in Statistics-Simulations and Computations, 39*(8), 1639 –1670. doi: 10.1080/03610918.2010.508862

Massy, W. F. (1965). Principal Components Regression in exploratory statistical research. *Journal of the American Statistical Association, 60*(309), 234-256. doi: 10.1080/01621459.1965.10480787

McDonald, G. C. & Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association, 70*(350), 407-416. doi: 10.1080/01621459.1975.10479882

Muniz, G. & Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics-Simulation and Computation, 38*(3), 621-630. doi: 10.1080/03610910802592838

Muniz, G., Kibria, B. M. G., Mansson, K. & Shukur, G. (2012). On Developing Ridge Regression Parameters: A Graphical Investigation. *SORT***.** *36*(2), 115-138.

Nordberg, L. (1982). A procedure for determination of a good ridge parameter in linear regression. *Communications in Statistics - Simulation and Computation, 11*(3), 285-309. doi: 10.1080/03610918208812264

Wichern, D. & Churchill, G. (1978). A comparison of ridge estimators. *Technometrics, 20*(3), 301–311. doi: 10.1080/00401706.1978.10489675

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiaah (Ed.). *Multivariate Analysis.* (pp.391-420) New York: Academic Press.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 452-460.
doi: 10.22237/jmasm/1493598300

ISSN 1538 − 9472

Control Charts for Mean for Non-Normally Correlated Data

J. R. Singh
Vikram University
Ujjain, India

Ab Latif Dar
Vikram University
Ujjain, India

Traditionally, quality control methodology is based on the assumption that serially-generated data are independent and normally distributed. On the basis of these assumptions the operating characteristic (OC) function of the control chart is derived after setting the control limits. But in practice, many of the basic industrial variables do not satisfy both the assumptions and hence one may doubt the validity of the inferences drawn from the control charts. In this paper the power of the control chart for the mean is examined when both the assumptions of independence and normality are not tenable. The OC function is calculated and compared with the normal population.

Keywords: Control chart, correlation, Edgeworth Series, standardized cumulants

Introduction

The quality control techniques currently used in industry are aimed at the detection of changes in the production process that result in quality defects. Quality control charts are currently the most widely-adopted control technique. Traditionally, quality control methodology is based on the assumption that serially-generated data are independent and normally distributed. Under these conditions, appropriate control limits for $\bar{X}$ can be worked out from the tables available in standard textbooks on statistical quality control. But in practice, many of the basic industrial processes do not satisfy both the assumptions and hence one may doubt the validity of the inference drawn from the control charts.

Alwan (1992) studied the effect of auto-correlation on control chart performance. Maragah and Woodall (1992) studied the effect of auto-correlation on the retrospective *X*-chart. Alwan and Roberts (1995) conducted investigations of control charts when the assumptions of normality, independence, or both are violated. Dar and Singh (2015) studied the effect of correlation on the power of

Ab Latif Dar is a Professor in the School of Studies in Statistics. Email them at: lateefdar.2007@rediffmail.com.

the $\bar{X}$ chart. The purpose of this study is to consider the power of the control chart and the effect of correlation on Type-I error and the OC function, and also to consider relaxing the assumption of normality and considering the production process to follow a non-normal distribution represented by the first four terms of an Edgeworth series.

Effect of Correlation on OC Function for Normal Case

Suppose that the observations $x_1, x_2, \ldots, x_n$ have a multivariate normal distribution with $E(x_i) = \mu$, $V(x_i) = \sigma^2$ and ρ is the common correlation coefficient between any x_i and x_j, $i \neq j$. Then

$$\begin{aligned} E(\bar{x}) &= \mu \\ \mathrm{Var}(\bar{x}) &= \frac{\sigma^2}{n}\left[1+(n-1)\rho\right] \\ &= \frac{\sigma^2}{n}T^2 \end{aligned} \tag{1}$$

where

$$T^2 = \left[1+(n-1)\rho\right] \tag{2}$$

The power of the control chart is judged by its OC function. The control chart for the mean is set up by drawing the central line at the process average θ and the control limits at

$$\theta \pm \frac{k\sigma}{\sqrt{n}}$$

where σ is the process standard deviation and n is the sample size. The OC function gives the probability that the control chart indicates the value θ as the process average, when it is actually not θ, but

$$\theta' = \theta + \frac{\gamma\sigma T}{\sqrt{n}}$$

where T^2 is as defined in equation (2).

The OC function is derived by integrating the distribution of the mean with θ' as the process average between the limits of the control chart.

For the normal population under correlated data,

$$\mathrm{f}\left(X\right)=\frac{1}{\sigma}\left[\phi\left(\frac{X-\theta}{\sigma}\right)\right] \tag{3}$$

The distribution of the sample mean is given by

$$\mathrm{g}\left(\bar{X}\right)=\frac{\sqrt{n}}{\sigma T}\left[\phi\left(\frac{\bar{X}-\theta}{\sigma T/\sqrt{T}}\right)\right] \tag{4}$$

where

$$\phi\left(t\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right) \quad \text{and} \quad \phi^r\left(t\right)=\frac{d^r}{dt^r}\phi\left(t\right)$$

The OC function is obtained after replacing θ in (4) by θ' and integrating it between the limits of the control chart as

$$L_N=\frac{\sqrt{n}}{\sigma T}\int_{\theta-\frac{k\sigma}{\sqrt{n}}}^{\theta+\frac{k\sigma}{\sqrt{n}}}\phi\left[\frac{\bar{X}-\left(\theta+\frac{\gamma\sigma T}{\sqrt{n}}\right)}{\frac{\sigma T}{\sqrt{n}}}\right]d\bar{X} \tag{5}$$

$$L_N=\frac{\sqrt{n}}{\sigma T}\int_{\theta-\frac{k\sigma}{\sqrt{n}}}^{\theta+\frac{k\sigma}{\sqrt{n}}}\phi\left[\frac{\bar{X}-\theta}{\sigma T/\sqrt{n}}+\gamma\right]d\bar{X} \tag{6}$$

Making the transformation

$$\frac{\bar{X}-\theta}{\sigma T/\sqrt{n}}=y$$

and $y - \gamma = t$ sequentially, the above integral simplifies to

$$L_N = \Phi\left(\frac{k}{T} + \gamma\right) + \Phi\left(\frac{k}{T} - \gamma\right) - 1 \tag{7}$$

The error of Type I gives the probability of searching for assignable causes when in fact there are no such causes. It is given by

$$\alpha = 1 - \int_{\theta - \frac{k\sigma}{\sqrt{n}}}^{\theta + \frac{k\sigma}{\sqrt{n}}} \mathrm{g}\left(\bar{X}\right) d\bar{X} \tag{8}$$

After integrating above as in the case of the OC function we will get

$$\alpha = 2\Phi\left(-\frac{k}{T}\right) \tag{9}$$

The Effect of Non-Normally Correlated Data on OC Function

For non-normal populations represented by the first four terms of an Edgeworth series,

$$\mathrm{f}\left(X\right) = \frac{1}{\sigma}\left[\begin{array}{c} \phi\left(\frac{\bar{X} - \theta}{\sigma}\right) - \frac{\lambda_3}{6}\phi^{(3)}\left(\frac{\bar{X} - \theta}{\sigma}\right) + \frac{\lambda_4}{24}\phi^{(4)}\left(\frac{\bar{X} - \theta}{\sigma}\right) \\ + \frac{\lambda_3}{72}\phi^{(6)}\left(\frac{\bar{X} - \theta}{\sigma}\right) \end{array}\right] \tag{10}$$

where $\lambda_3 = \sqrt{\beta_1}$ and $\lambda_4 = \left(\beta_2 - 3\right)$ are the standardized third and fourth cumulants, respectively.

The distribution of the sample mean for correlated data can be derived, by following Gayen (1952), as

$$g\left(\bar{X}\right)=\frac{\sqrt{n}}{\sigma T}\left[\begin{array}{l}\phi\left(\frac{\bar{X}-\theta}{\sigma T/\sqrt{n}}\right)-\frac{\lambda_3 T}{6\sqrt{n}}\phi^{(3)}\left(\frac{\bar{X}-\theta}{\sigma T/\sqrt{n}}\right)\\ \quad+\frac{\lambda_4 T^2}{24n}\phi^{(4)}\left(\frac{\bar{X}-\theta}{\sigma T/\sqrt{n}}\right)+\frac{\lambda_3^2 T^2}{72n}\phi^{(6)}\left(\frac{\bar{X}-\theta}{\sigma T/\sqrt{n}}\right)\end{array}\right] \tag{11}$$

The OC function is obtained after replacing θ in equation (11) by θ' and integrating it between the limits of the control chart, i.e. between $\theta\pm\left(k\sigma\right)/\sqrt{n}$. Integrating in the similar way as for the normal case, we get

$$L'=L_N-L'_u+L'_b \tag{12}$$

where L_N is given by equation (7). The other two terms of the OC function are given by

$$L'_u=\frac{T^2}{72n}\left[\frac{12\lambda_3\sqrt{n}}{T}\phi^{(2)}\left(\frac{k}{T}-\gamma\right)-3\lambda_4\phi^{(3)}\left(\frac{k}{T}-\gamma\right)-\lambda_3^2\phi^{(5)}\left(\frac{k}{T}-\gamma\right)\right] \tag{13}$$

$$L'_b=\frac{T^2}{72n}\left[\frac{12\lambda_3\sqrt{n}}{T}\phi^{(2)}\left(\frac{k}{T}+\gamma\right)+3\lambda_4\phi^{(3)}\left(\frac{k}{T}+\gamma\right)+\lambda_3^2\phi^{(5)}\left(\frac{k}{T}+\gamma\right)\right] \tag{14}$$

The Type-I error for the non-normal population works out to be

$$\alpha'=1-\int_{\theta-\frac{k\sigma}{\sqrt{n}}}^{\theta+\frac{k\sigma}{\sqrt{n}}}g\left(\bar{X}\right)d\bar{X}=\alpha-c \tag{15}$$

where α as defined by equation (9) is the Type-I error when the population is normal and dependent, and

$$c=\frac{T^2}{36n}\left[3\lambda_4\phi^{(3)}\left(\frac{k}{T}\right)+\lambda_3^2\phi^{(5)}\left(\frac{k}{T}\right)\right] \tag{16}$$

is the correction for non-normality and dependencies in Type-I error.

Table 1. Value of Type-I error for normally and correlated data

	K = 2				*K* = 3			
n	**ρ = 0.0**	**ρ = 0.2**	**ρ = 0.5**	**ρ = 0.8**	**ρ = 0.0**	**ρ = 0.2**	**ρ = 0.5**	**ρ = 0.8**
5	0.04550	0.13603	0.24821	0.32911	0.00269	0.02534	0.08326	0.14323
10	0.04550	0.23200	0.39377	0.48491	0.00270	0.07300	0.20083	0.29480
15	0.04550	0.30490	0.47950	0.56692	0.00270	0.12381	0.28884	0.39040

Table 2. Value of OC function for normally and correlated data

		K = 2				*K* = 3			
n	***γ***	**ρ = 0.0**	**ρ = 0.2**	**ρ = 0.5**	**ρ = 0.8**	**ρ = 0.0**	**ρ = 0.2**	**ρ = 0.5**	**ρ = 0.8**
5	-2	0.9932	0.3050	0.1981	0.1514	0.8413	0.5932	0.3942	0.2956
	-1	0.9997	0.6818	0.5458	0.4663	0.9772	0.8911	0.7647	0.6717
	0	0.9999	0.8639	0.7517	0.6708	0.9973	0.9746	0.9167	0.8567
	1	0.9997	0.6818	0.5458	0.4663	0.9772	0.8911	0.7647	0.6717
	2	0.9932	0.3050	0.1981	0.1514	0.8413	0.5932	0.3942	0.2956
10	-2	0.5000	0.2098	0.1235	0.0930	0.8413	0.4179	0.2350	0.1693
	-1	0.8400	0.5633	0.4095	0.3368	0.9772	0.7835	0.5986	0.4987
	0	0.9545	0.7680	0.6062	0.5151	0.9973	0.9270	0.7992	0.7052
	1	0.8400	0.5633	0.4095	0.3368	0.9772	0.7835	0.5986	0.4987
	2	0.5000	0.2098	0.1235	0.0930	0.8413	0.4179	0.2350	0.1693
15	-2	0.5000	0.1638	0.0946	0.0717	0.8413	0.3222	0.1727	0.1248
	-1	0.8400	0.4890	0.3409	0.2766	0.9772	0.6995	0.5045	0.4124
	0	0.9545	0.6951	0.5205	0.4331	0.9973	0.8762	0.7112	0.6096
	1	0.8400	0.4890	0.3409	0.2766	0.9772	0.6995	0.5045	0.4124
	2	0.5000	0.1638	0.0946	0.0717	0.8413	0.3222	0.1727	0.1248

Results and Conclusion

For normal populations with correlation coefficient ρ = 0, 0.2, 0.5, and 0.8, the values of Type-I error have been computed and given in Table 1 for k = 2, 3 and n = 5, 10, 15. Table 1 clearly indicates that the effect of correlation on Type-I error is quite substantial as the error increases with the increase in ρ. For example, for n = 5, k = 2, and ρ = 0, 0.2, 0.5, and 0.8, the corresponding values of Type-I error are 0.04550, 0.13605, 0.24821, and 0.32911. Though the effect goes on decreasing with increasing k, it still affects the value of Type-I error quite largely. For non-normal populations we have a similar result (Table 3) as the error goes on increasing with an increase in the value of ρ, λ_3, and λ_4. From Table 2, it is evident that the value of the OC are affected seriously as the correlation between the observations increases. For example, for ρ = 0, k = 2, n = 5, and $\gamma = \pm 1$, the value

of the OC is 0.9997, while for $\rho = 0.2, 0.5, 0.8$, $k = 2$, and $n = 5$, the value reduces to 0.6818, 0.5458, 0.4663. For other values of $n = 10, 15$, we have a similar results. The values of the OC for non-normal populations with $k = 2$, $n = 5$ and for different values of $\rho = 0, 0.2, 0.5, 0.8$ are given in Table 4. For $\rho = 0$ and $(\lambda_3, \lambda_4) = (0, 0)$ we get tabulated values of Singh, Sankle, and Ahmad (2012), which are shown in Table 4. The effect of correlation on the OC function remains more or less of the same magnitude when we move from normal to non-normal populations. As is evident from the Table 4, for $\rho = 0$, $\lambda_3 = 0$, $\lambda_4 = 0$, and $\gamma = \pm 1$, the value of the OC function is 0.8400 while for $\rho = 0.5$, $\lambda_3 = 0.5$, $\lambda_4 = 0.5$, and $\gamma = \pm 1$, the corresponding value of the OC function is reduced to 0.3725. On changing λ_3 (skewness), λ_4 (kurtosis), or both at the same time, the value of the OC is affected. Therefore, it may be inferred that the violation in the assumptions of independence and normality have a serious effect on the control chart performance and it is advisable to take into account the dependence and non-normality of the parent population while designing control charts.

Table 3. Values of the Type-I error for non-normally correlated data

			***K* = 2**				***K* = 3**			
ρ	***n***	λ_3	λ_4**=0.0**	λ_4**=0.5**	λ_4**=1.0**	λ_4**=2.0**	λ_4**=0.0**	λ_4**=0.2**	λ_4**=0.5**	λ_4**=0.8**
0.0	5	0.0	0.0455	0.0464	0.0473	0.0491	0.0027	0.0034	0.0040	0.0054
		0.5	0.0442	0.0451	0.0459	0.0477	0.0028	0.0035	0.0041	0.0055
	10	0.0	0.0455	0.0459	0.0464	0.0473	0.0027	0.0030	0.0034	0.0040
		0.5	0.0448	0.0453	0.0457	0.0466	0.0028	0.0031	0.0034	0.0041
	15	0.0	0.0455	0.0458	0.0461	0.0467	0.0027	0.0029	0.0031	0.0036
		0.5	0.0451	0.0454	0.0456	0.0462	0.0027	0.0030	0.0032	0.0036
0.2	5	0.0	0.1360	0.1338	0.1315	0.1269	0.0253	0.0275	0.0297	0.0341
		0.5	0.1349	0.1326	0.1303	0.1258	0.0235	0.0257	0.0279	0.0323
	10	0.0	0.2320	0.2277	0.2234	0.2149	0.0730	0.0734	0.0737	0.0744
		0.5	0.2332	0.2290	0.2247	0.2161	0.0711	0.0715	0.0718	0.0725
	15	0.0	0.3049	0.2999	0.2950	0.2850	0.1238	0.1226	0.1213	0.1188
		0.5	0.3073	0.3023	0.2973	0.2874	0.1228	0.1215	0.1203	0.1178
0.5	5	0.0	0.2482	0.2384	0.2285	0.2088	0.0833	0.0833	0.0833	0.0833
		0.5	0.2516	0.2418	0.2319	0.2122	0.0794	0.0794	0.0794	0.0794
	10	0.0	0.3938	0.3814	0.3691	0.3445	0.2008	0.1938	0.1867	0.1727
		0.5	0.4012	0.3889	0.3766	0.3519	0.2020	0.1949	0.1879	0.1738
	15	0.0	0.4795	0.4673	0.4551	0.4307	0.2888	0.2788	0.2687	0.2487
		0.5	0.4878	0.4756	0.4634	0.4390	0.2933	0.2833	0.2732	0.2531

Table 3, continued.

			$K = 2$				$K = 3$			
ρ	n	λ_3	λ_4=0.0	λ_4=0.5	λ_4=1.0	λ_4=2.0	λ_4=0.0	λ_4=0.2	λ_4=0.5	λ_4=0.8
0.8	5	0.0	0.3291	0.3118	0.2944	0.2598	0.1432	0.1372	0.1312	0.1192
		0.5	0.3381	0.3208	0.3034	0.2688	0.1411	0.1351	0.1291	0.1171
	10	0.0	0.4849	0.4662	0.4474	0.4099	0.2948	0.2791	0.2634	0.2320
		0.5	0.4978	0.4790	0.4603	0.4228	0.3020	0.2863	0.2706	0.2392
	15	0.0	0.5669	0.5494	0.5318	0.4967	0.3904	0.3722	0.3541	0.3177
		0.5	0.5799	0.5623	0.5447	0.5096	0.4013	0.3832	0.3650	0.3286

Table 4. Values of OC function for non-normally correlated data

		(λ_3, λ_4)							
ρ	γ	(0.0,0.0)	(0.0,0.5)	(0.0,1.0)	(0.0,2.0)	(0.5,0.0)	(0.5,0.5)	(0.5,1.0)	(0.5,2.0)
0.0	-2	0.5000	0.5000	0.4999	0.4999	0.4850	0.4850	0.4849	0.4849
	-1	0.8400	0.8408	0.8417	0.8434	0.8396	0.8413	0.8430	0.8464
	0	0.9545	0.9540	0.9536	0.9527	0.9545	0.9536	0.9527	0.9509
	1	0.8400	0.8408	0.8417	0.8434	0.8404	0.8420	0.8437	0.8471
	2	0.5000	0.5000	0.4999	0.4999	0.5150	0.5149	0.5149	0.5148
0.2	-2	0.2098	0.2063	0.2028	0.1957	0.2022	0.1986	0.1951	0.1881
	-1	0.5633	0.5638	0.5643	0.5653	0.5633	0.5430	0.5435	0.5445
	0	0.7680	0.7723	0.7766	0.7851	0.7680	0.7723	0.7766	0.7851
	1	0.5633	0.5638	0.5643	0.5653	0.5633	0.5846	0.5851	0.5861
	2	0.2098	0.2063	0.2028	0.1957	0.1229	0.2139	0.2104	0.2033
0.5	-2	0.1235	0.1178	0.1120	0.1006	0.1229	0.1172	0.1115	0.1000
	-1	0.4095	0.4069	0.4042	0.3990	0.3752	0.3725	0.3699	0.3646
	0	0.6062	0.6186	0.6309	0.6555	0.6062	0.6186	0.6309	0.6555
	1	0.4095	0.4069	0.4042	0.3990	0.4439	0.4413	0.4386	0.4333
	2	0.1235	0.1178	0.1120	0.1006	0.1241	0.1183	0.1126	0.1012
0.8	-2	0.0930	0.0860	0.0790	0.0649	0.0968	0.0898	0.0828	0.0687
	-1	0.3368	0.3314	0.3260	0.3152	0.3314	0.3314	0.3260	0.3152
	0	0.5151	0.5338	0.5526	0.5901	0.5338	0.5338	0.5526	0.5901
	1	0.3368	0.3314	0.3260	0.3152	0.3314	0.3314	0.3260	0.3152
	2	0.0930	0.0860	0.0790	0.0649	0.0860	0.0860	0.0790	0.0649

References

Alwan, L. C. (1992). Effects of autocorrelation on control chart performance. *Communications in Statistics – Theory and Methods, 21*(4), 1025-1049. doi: 10.1080/03610929208830829

Alwan, L. C., & Roberts, H. V. (1995). The problem of misplaced control limits. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 44*(3), 269-278. doi: 10.2307/2986036

Dar, A. L., & Singh, J. R. (2015). The power of $\bar{X}$-chart in presence of data correlation. *Journal of Reliability and Statistical Studies, 8*(1), 25-30. Retrieved from http://www.jrss.in.net/assets/8103.pdf

Gayen, A. K. (1952). On setting up control charts for non-normal samples. *Indian Society for Quality Control Bulletin, 53*, 43-47.

Maragah, H. D, & Woodall, W. H. (1992). The effect of autocorrelation on the retrospective *X*-chart. *Journal of Statistical Computation and Simulation, 40*(1-2), 29-42. doi: 10.1080/00949659208811363

Singh, J. R., Sankle, R., & Ahmad, M. (2012). Control charts for mean under correlated data. *Journal of Rajasthan Statistical Association, 1*(1), 21-30.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 461-480.
doi: 10.22237/jmasm/1493598420

ISSN 1538 − 9472

Plant Leaf Image Detection Method Using a Midpoint Circle Algorithm for Shape-Based Feature Extraction

B. Vijaya Lakshmi
K.L.N. College of Engineering
Tamil Nadu, India

V. Mohan
Thiagarajar College of Engineering
Tamil Nadu, India

Shape-based feature extraction in content-based image retrieval is an important research area at present. An algorithm is presented, based on shape features, to enhance the set of features useful in a leaf identification system.

Keywords: Centroid, edge detection, Euclidean distance, feature extraction, midpoint circle

Introduction

The recognition and identification of plants permits exploration of the genetic relationship and evolutionary law of plant systems. When recognizing and identifying plants, leaf, flower, stem, fruit and other discriminating features are observed. Eventually, computers will conduct the recognition of plants automatically or semi-automatically.

Content-based retrieval methods may be classified by features such as shape, color, or texture. They are divided into subclasses by the types of algorithm used for constructing the feature vector.

Shape is an important visual feature, and it is one of the primary features for image content description. However, shape description is difficult, because it is demanding to define perceptual shape features that measure the similarity between the shapes. The problem is more complex if shape is corrupted with noise, defection, arbitrary distortion, or occlusion. Shape has been an active research area for over thirty years. In the past, shape research has been driven

B. Vijaya Lakshmi is an Assistant Professor in the Department of Master of Computer Applications. Email her at: bviji0677@gmail.com. Dr. V. Mohan is Professor and Head of the Department of Mathematics. Email him at vmohan@tce.edu.

mainly by object recognition. As a result, techniques of shape representation and description mostly target particular applications such as leaf classification.

The centroid radius model is frequently used as a shape feature. The centroid radius describes the length of a radius from its centroid to its boundary, and the model captures these lengths at regular intervals as shape descriptors, using the Euclidean distance. These distances are considered features for shape description.

Let θ be the regular interval (measured in degrees) between the radii. At that point, the number of intervals given by $k = (360/\theta)$. Numbers of features depend on the fixed value of θ, and this strategy cannot measure many features of the specific region. This conventional radii model generates the vector that is the normalized length of radius for shape representation. The vector depends on the order of the radii.

The purpose of this study is to propose a new shape descriptor, based on a center point and a border of the circle and contour of the leaf, which is more effective for shape description and retrieval. It overcomes the limitation of earlier descriptors by calculating radius (distance) in distinctive ways, and by upgrading the set of feature values. Distance will be calculated between the contour of the circle and contour of leaf image, rather than between the center point and border of leaves.

This descriptor leads to identification of plant leaves based on a leaf query using shape features such as contour of leaf, center point of the leaf, and border of the circle in addition to radius. The prototype of this system has been implemented and the experiment results prove the effectiveness and superiority of the proposed method.

Existing Method

Centroid Radii Model (CRM)

Tan et al. (2003) proposed the centroid radii model (CRM) for estimating shapes of objects in images. A shape is defined to be an area of black with a background of white. Each pixel is represented by its color (black or white) and its x/y coordinates on the canvas. The boundary of a shape consists of a series of boundary points. A boundary point is a black pixel with at least one white pixel as its neighbor. Let (x_i, y_i), $i = 1, \ldots, n$ represent the shape having n boundary points. The centroid is located at the position $C(X_c,Y_c)$ which are respectively, the average of the x and y co-ordinates for all black pixels:

$$X_c = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

$$Y_c = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2}$$

A radius is a straight line joining the centroid to a boundary point. In the CRM, lengths of a shape's radii from its centroid to the boundary are captured as the shape descriptor at regular intervals using the Euclidean distance. More formally, let θ be the regular interval (measured in degrees) between radii (Figure 1). Then, the number of intervals is given by $k = (360/\theta)$. The length L_i of the i^{th} radius formed by joining the centroid $C(X_c,Y_c)$ to the i^{th} sample point $s_i(X_i,Y_i)$ is given by:

$$L_i = \sqrt{\left(X_c - x_i\right)^2 + \left(Y_c - y_i\right)^2} \tag{3}$$

All radii lengths are normalized by dividing with the longest radius length from the set of radii lengths extracted. Let the individual radii lengths be $L_1, L_2, L_3, \ldots, L_k$ Where k is the total number of radii drawn at an angle. If the maximum radius length is $L_{\max}$, the normalized radii lengths are given by:

$$l_i = \frac{L_i}{L_{\max}},\ i = 1,\ldots,k \tag{4}$$

Furthermore, without loss of generality, suppose that the intervals are taken clockwise starting from the x-axis direction (0^0). Then, the shape descriptor can be represented as a vector consisting of an ordered sequence of normalized radii lengths:

$$S = \left\{l_0, l_\theta, l_{2\theta}, \ldots, l_{(k-1)\theta}\right\} \tag{5}$$

Here, $l_{i\theta}, 0 \le i \le (K-1)$ is the $(i+1)^{\text{th}}$ radius from the centroid to the boundary of the shape. With sufficient number of radii, dissimilar shapes can be differentiated from each other.

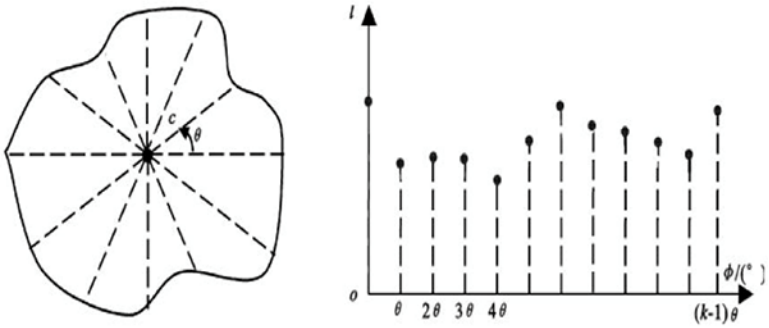


Figure 1. The centroid radii modeling of shape

Midpoint Circle Algorithm

One of the fundamental graphics primitives is the circle. The most efficient conventional algorithm for drawing a circle is the mid-point algorithm, based on Bresenham's approach (Bresenham, 1977; Ray, 2006). The mid-point algorithm starts on a quadrant boundary, and generates pixel by pixel, making either an axial or a diagonal move, depending on the sign of the decision variable.

A circle is defined as a set of points that are all at a given distance r from a center positioned at (X_c,Y_c). This is represented mathematically by the equation

$$(x-x_c)^2+(y-y_c)^2=r^2 \tag{6}$$

Using equation (6), calculate the value of y for each given value of x as

$$y=y_c\pm\sqrt{r^2-(X_c-X)^2} \tag{7}$$

Thus, it is possible to calculate different pairs by giving step increments to x and calculating the corresponding value of y. The midpoint circle algorithm uses an alternative approach, wherein the pixel positions along the circle are determined on the basis of incremental calculations of a decision parameter.

Let

$$f(x,y)=(x-x_c)^2+(y-y_c)^2-r^2 \tag{8}$$

Thus, $f(x,y) = 0$ represents the equation of a circle. Further, from coordinate geometry, for any point the following holds:

1. $f(x,y) = 0 \rightarrow$ The point lies on the circle.
2. $f(x,y) < 0 \rightarrow$ The point lies within the circle.
3. $f(x,y) > 0 \rightarrow$ The point lies outside the circle.

In the midpoint circle algorithm, the decision parameter at the k^{th} step is the circle function evaluated using the coordinates of the midpoint of the two pixel centers, which are the next possible pixel position to be plotted.

Assume that unit increments to x in the plotting process given, and the y position is determined using this algorithm. Assuming that the k^{th} pixel is plotted at (X_k, Y_k) to determine whether the pixel at the position $(X_k + 1, Y_k)$, or the one at $(X_k + 1, Y_k - 1)$, is closer to the circle as shown in Figure 2.

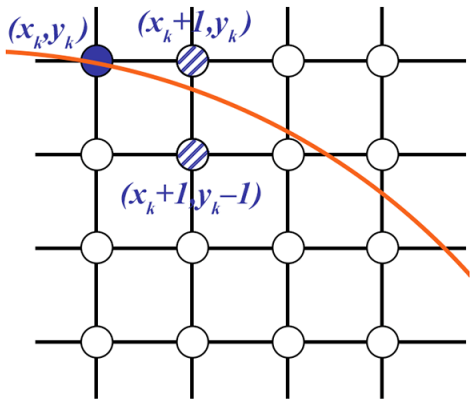


Figure 2. Plotting of the midpoint

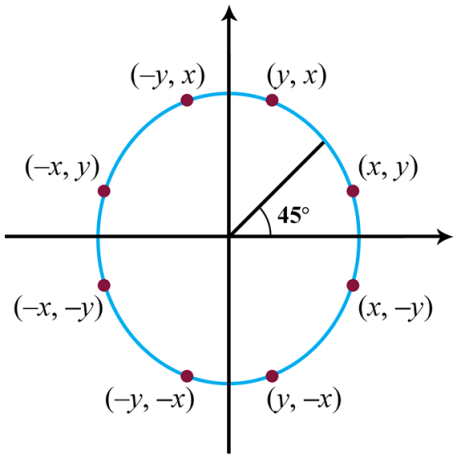


Figure 3. Eight-way symmetry

The decision parameter p_k at the k^{th} step is the circle function evaluated at the midpoint of these two pixels.

The coordinates of the midpoint of these two pixels are $(X_k + 1, Y_k - ½)$. Thus p_k,

$$p_k = f\left(X_k + 1, Y_k - \tfrac{1}{2}\right) = X_k + 1^2 + Y_k - \tfrac{1}{2}^2 - r^2 \qquad (9)$$

Symmetry of a circle calculation of a circle point (x,y) in one octant yields the circle points shown for the other seven octants, as shown in Figure 3.

The steps involved in the midpoint circle algorithm are as follows:

1. Input radius r and circle center (X_c,Y_c), and obtain the first point on the circumference of a circle centered on the origin as

$$(X_0, Y_0) = (0, r) \tag{10}$$

2. Calculate the initial value of the decision parameter as

$$p_0 = \frac{5}{4} - r \tag{11}$$

3. At each X_k position, starting at $k = 0$, perform the following test: if $p_k < 0$, the next point along the circle centered on (0,0) is (X_{k+1},Y_k) and p

$$p_{k+1} = p_k + 2x_{k+1} + 1 \tag{12}$$

otherwise, the next point along the circle is $(X_k + 1, Y_k - 1)$. And

$$p_{k+1} = p_k + 2x_{k+1} + 1 - 2y_{k+1} \tag{13}$$

where

$$2x_{k+1} = 2x_k - 2 \text{ and } 2y_{k+1} = 2y_k - 2.$$

4. Determine symmetry points in the other seven octants.
5. Move each calculated pixel position (x,y) onto the circular path centered on (X_c,Y_c) and plot the coordinate values.

$$x = x + x_c, y = y + y_c$$

6. Repeat steps 3 through 5 until $x \geq y$.

By using this algorithm, a circle will be formed in a two-dimensional plane.

Related Works

Chaki and Parekh (2011) proposed an approach that consists of comparing binary versions of the leaf images through superposition and using the sum of nonzero pixel values of the resultant as the feature vector. Consider the invariant (M-I) model and centroid radii (C-R) model. In M-I, central moments and normalized central moments are calculated. In C-R, the length of a shape's radii from its centroid of the boundary is captured at regular intervals as the shape descriptor using Euclidean distance. In the method, two images are in binary superposition; a large-value sum would indicate high similarity between images, and a small-value sum would indicate low similarity between images. A comparison of the recognition accuracy of these two methods with a proposed method used 180 leaf images from the plantScan database, three classes of 60 samples each. From that set, 120 images were used for training and 60 images were tested.

Lee and Hong (2013) proposed a leaf recognition system for plant classification, employing major vein, frequency domain data by using fast Fourier transformation. Dilation and erosion operations were used to extract leaf veins, and a projection histogram was calculated for both horizontal and vertical directions in order to measure the vein distribution of the leaf. The authors extracted 10 features, using fast Fourier transformation, distance, and phase; another 10 features were extracted using leaf length, width, area, and perimeter; and a final feature, convex hull, was also extracted. This system was applied to 1970 leaf images, consisting of 32 types of leaves with 50 to 77 samples each and implemented in VC++ 6.0, Intel OpenCV Library. The accuracy of this method was reported at 97.19%.

Deokar, Zope, and Suralkar (2013) presented feature point extraction. The feature points were extracted from a leaf image based on the geometric center. The authors proposed two schemes, vertical and horizontal, to extract feature points. In the vertical scheme, leaf images are split vertically in half with respect to a central point. Each half is then split horizontally in a number of repetitions, until 14 feature points are obtained. This process is repeated starting on the leaf's horizontal access, obtaining 14 more feature points for a total of 28.

Fulsoundar, Kadlag, Bhadale, Bharvirkar & Godse (2014) created an Android application to identify plant species based on photographs of the plant's leaves taken with a mobile phone. At the heart of this application is an algorithm that acquires morphological features of the leaves, computes well-documented metrics consisting of the angle code histogram (ACH), and classifies the species. The first algorithm was prepared against several samples of known plant species

after being used to classify unknown query species. Supported by features designed into the application such as touch screen image rotation and contour preview, the algorithm has been very successful in properly classifying species inside the training library.

Bong, M. F., Sulong, G., Kumoi, R., & Rahim, M. S. M. (2015) proposed a novel approach to cluster the species of plants based on their lobes, sinuses and margin. Firstly, all the boundary points in a clockwise or anticlockwise direction were selected. Then, a center point for leaf boundary points was estimated, and used to compute the distance between the leaf boundary points and the center. Next, the peaks and valleys from the computed distance were located, where peaks represent lobes and valleys represent the sinuses. The number of peaks and valleys was calculated to cluster the plant, according to a rule-based method. The accuracy of this method for plant clustering is up to 100 percent.

The Proposed Circle-based Radii Model (CBRM)

A new Circle-based Radii Model is now proposed for shape descriptor. A circle is formed by using a midpoint circle algorithm based on the center point of the leaf, as discussed earlier under that heading. Here, the radius is 0.5, after forming the circle; the proposed method uses two phases to obtain four feature points on the circle of a leaf image in the two-dimensional plane.

Architecture of the Circle-based Radii Model

The architecture of proposed circle-based radii model for the plant leaf recognition of 2D objects is given in the Figure 4.

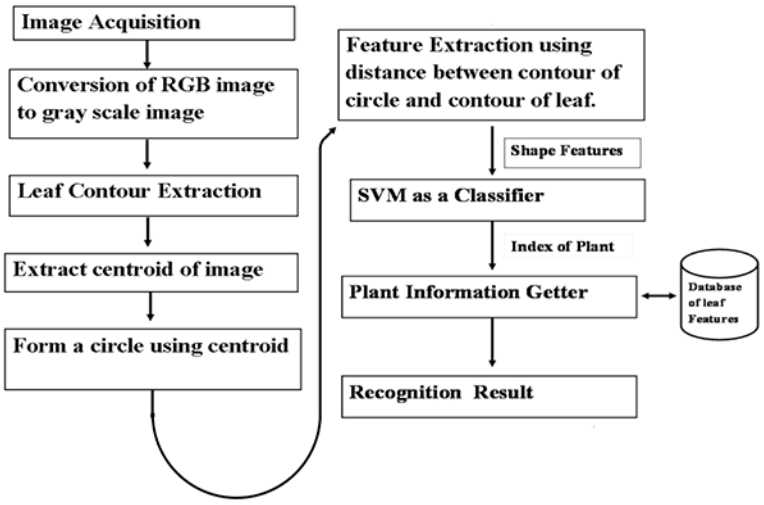


Figure 4. Architecture for leaf recognition system

It includes the following stages: preprocessing, circle formation and feature extraction.

Preprocessing

The leaf images may be acquired utilizing an advanced digital camera. There is no confinement on image resolution or format. Resulting digital images are usually in RGB color space; some may be grayscale. The fundamental objective of preprocessing is to a) identify the leaf in an image and b) discard all information other than the leaf shape. The initial step is to resize all the leaves to 256×256, then convert any RGB image to grayscale. Below is the equation used to convert RGB value of a pixel to its grayscale value.

$$Gray = 0.2989*R + 0.5870*G + 0.1140*B \quad (14)$$

where R, G, B corresponds to color of the pixel.

Finally, the Sobel edge detection method is applied to the resulting image. The entire procedure is shown in Figure 5.

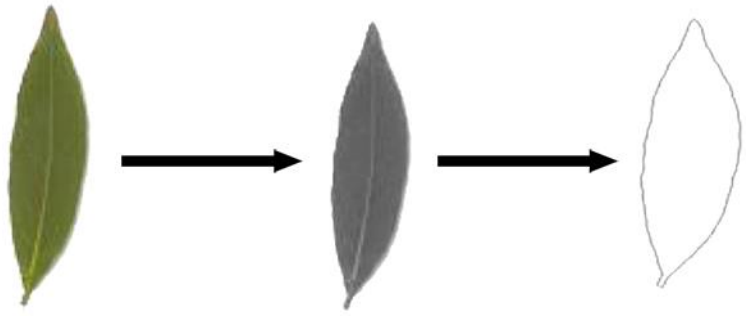


Figure 5. Sample of color to gray conversion and contour extraction of Leaf

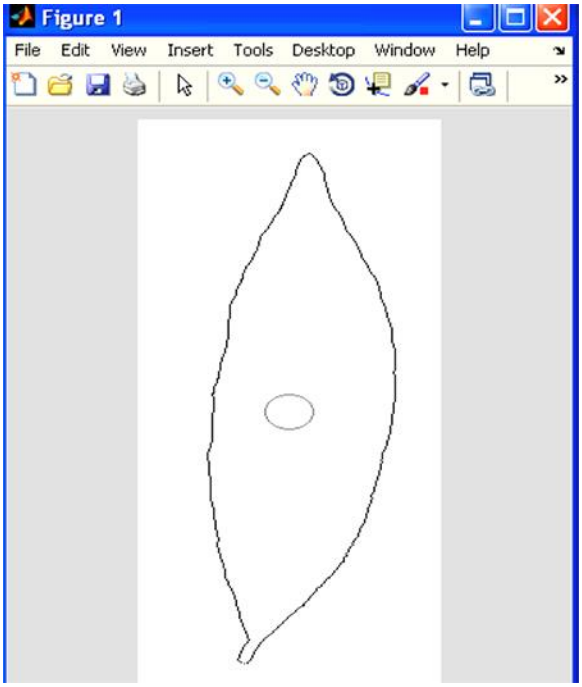


Figure 6. Circle formation within leaf

Circle Formation inside the Leaf

Once the contour of the leaf image is detected, a circle is formed based on the center point of the leaf image by using the midpoint circle algorithm. In the midpoint circle algorithm, an eight-way symmetry sample is used as in Figure 3.

The midpoint circle algorithm works on the same midpoint concept as Bresenham's line algorithm (Bresenham, 1977). Figure 6 shows the circle arrangement inside the leaf image.

Feature Extraction

The proposed techniques of the circle-based radii model (CBRM) are discussed here. Subsequent to forming the circle, the method obtains four feature points on the circle of a leaf image in the two-dimensional plane. The distance is calculated between the contour of the circle and the contour of the leaf image as shown in Figure 7. They are described as:

The distance between the contour of the circle (x_i,y_i) and the contour of the leaf image (x_r,y_r) is called *Realdistance*, and is shown in Figure 7.

The distance between a point in a vertical or horizontal line running through the center point (x_c,y_c) and contour of circle (x_i,y_i) is called *Imaginarydistance*.

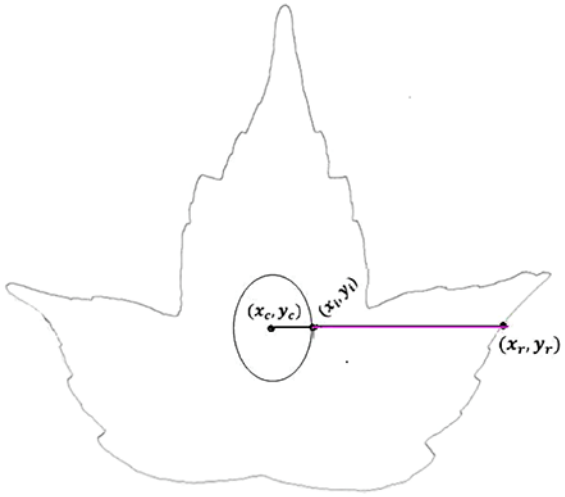


Figure 7. Circle-based radii model (CBRM)

In the proposed method, the *Realdistance* (RD) are calculated using equation (15) for feature extraction:

$$RD_i = \sqrt{\left(X_i - x_r\right)^2 + \left(Y_i - y_r\right)^2} \tag{15}$$

In the first phase, splitting the circle vertically through the center derives its top and bottom points. Then, left-to-right row-wise scanning from the top point of the circle to the leaf contour is conducted to the bottom point, as illustrated in Figure 8(a). Euclidean distance is calculated between the contour of the circle and contour of the leaf image, rather than between the center point and border of leaves. This set of *Realdistance* values is considered the first set of feature points for shape description. The process is repeated in the opposite direction, as shown in Figure 8(b), resulting in the second set of shape descriptor features.

The second phase repeats this process, splitting the circle horizontally and measuring similar distances vertically, to derive the third and fourth set of shape descriptor features, as illustrated in Figure 8(c) and 8(d).

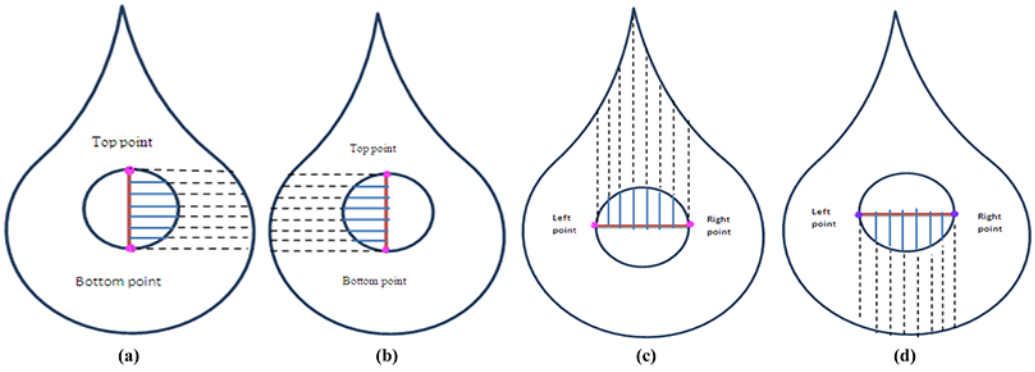


Figure 8. Real distance calculation (a) Right direction (b) Left direction (c) Top direction (d) Down direction

Algorithm: Circle-based Radii Model

The CBRM algorithm for radius calculation is summarized in the following steps.

Input: Image and radius
Output: *N* number of features

Step 1: Load the image and detect the border of the image.

Step 2: Assign image size as [*a*,*b*]

Step 3: Find the center point of the image (x_c,y_c)

Step 4: Draw the circle by using Midpoint circle algorithm.

Assign:

Center point	=	(x_c,y_c)
Top point	=	$(x_c - radius, y_c)$
Bottom point	=	$(x_c + radius, y_c)$
Leaf point	=	$(x_c,y_c - radius)$
Right Point	=	$(x_c,y_c + radius)$

Count = 0, Flag = 0

The following steps are used to extract the features by vertical splitting of the circle

Step 1: Divide the circle in half vertically from the top point to the bottom point.

Apply steps 2 to 7 to right half of the circle.

Step 2: Start scanning from top point to the bottom point in a row-wise manner.

Step 3: Take the every column value and check the following conditions one by one.

Step 4: Check image (i,j) is equal to the border of the circle or not. If it is bordered, assign Flag = 1.

Step 5: Check flag > 0 (zero), get the value of (x_i,y_i) coordinate

Step 6: Check image (i,j) is equal to border of the image. If it is

i. Get the value of (x_r,y_r) coordinate
ii. Calculate the distance using equation (15)
iii. Reassign Flag = 0 (zero)

Step 7: Go to the next row, repeat the steps from 4 to 7 until the bottom point of the circle. These are considered the first set of features for shape description

Step 8: Repeat the steps 2 to 7 to the left half side of the circle. This process will produce the second set of feature values.

The following steps are used to extract sets three and four:

Step 10: Divide the circle in half horizontally from the leftmost point to the rightmost point.

Step 11: Start reading from the left point to the right point in column wise manner.

Step 12: Take the every row value of each column and check the following conditions one by one, as in steps 4 to 6.

Step 13: Go to the next column, repeat the steps 4 to 6 until right point of the circle.

Step 14: Repeat the steps 4 to 6 to bottom half side of the circle.

Calculating all distances in four directions in this way, the number of features and sets of feature values will be obtained.

Experiments and Results

Plant leaves classification framework is developed and executed in MATLABR2009b, to test the retrieval effectiveness and performance of the CBRM. To evaluate the effectiveness of the proposed approach, recognition procedure is carried out on a large texture dataset, provided by the Intelligent Computing Laboratory at the Chinese Academy of Science (ICL CAS, n.d.). Samples of leaf images belonging to various classes are shown in Figure 9. A sample converted gray scale leaf image is shown in Figure 10.

Experiments are performed by using 50 of 220 classes from the ICL dataset. Fifteen leaves are selected randomly from each class. 750 total (15×50 classes) sample leaves are trained. In conventional centroid radii model, twelve features

are extracted when θ = 30°. Using the proposed method, 11 features are extracted from each direction, for a total of 44. Extracted feature values are stored in feature vector. In the end, the dataset generates 33,000 (750 leaves × 44 features) feature values.

Figure 9. A sample of plant leaf images taken from the ICL plant leaf dataset

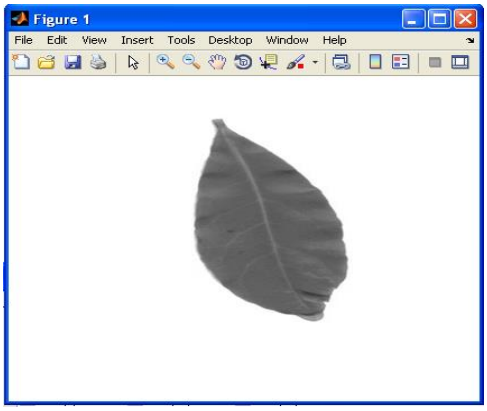
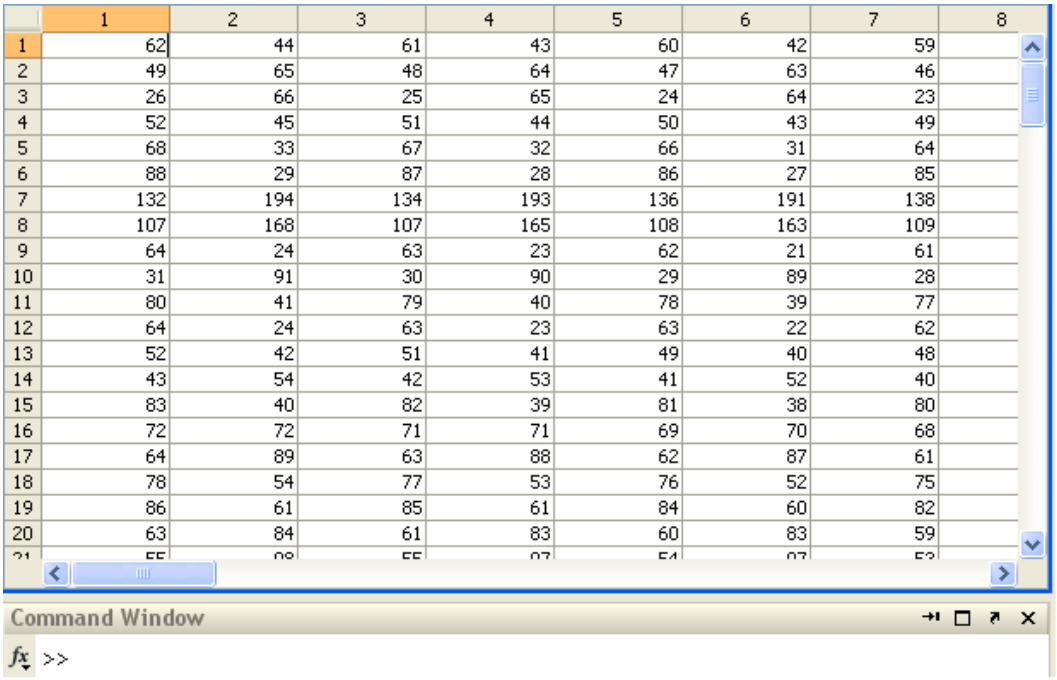


Figure 10. Transformed gray scale leaf image

Presented in Figure 11 are the results obtained with CBRM. In the testing phase, 15,000 leaves were tested randomly for the accuracy of the proposed technique. The recognition process is carried out by comparing the length of the tested circle radii with the reference. The input query image feature vector is compared with trained feature set using SVM classifier. Average of recognition of each class is given in Table 1. The performance comparison of the proposed CBRM with the centroid radii model is shown in Figure 12. The accuracy rate of the proposed method is shown to be 93.33%.



	1	2	3	4	5	6	7	8
1	62	44	61	43	60	42	59	
2	49	65	48	64	47	63	46	
3	26	66	25	65	24	64	23	
4	52	45	51	44	50	43	49	
5	68	33	67	32	66	31	64	
6	88	29	87	28	86	27	85	
7	132	194	134	193	136	191	138	
8	107	168	107	165	108	163	109	
9	64	24	63	23	62	21	61	
10	31	91	30	90	29	89	28	
11	80	41	79	40	78	39	77	
12	64	24	63	23	63	22	62	
13	52	42	51	41	49	40	48	
14	43	54	42	53	41	52	40	
15	83	40	82	39	81	38	80	
16	72	72	71	71	69	70	68	
17	64	89	63	88	62	87	61	
18	78	54	77	53	76	52	75	
19	86	61	85	61	84	60	82	
20	63	84	61	83	60	83	59	

Figure 11. Set of feature value of trained leaves

Table 1. CBRM recognition accuracy for each class of leaf

Class	Accuracy	Class	Accuracy	Class	Accuracy	Class	Accuracy	Class	Accuracy
1	90%	*11*	95%	*21*	89%	*31*	94%	*41*	88%
2	92%	*12*	88%	*22*	89%	*32*	93%	*42*	87%
3	94%	*13*	84%	*23*	92%	*33*	95%	*43*	96%
4	90%	*14*	94%	*24*	92%	*34*	94%	*44*	95%
5	90%	*15*	93%	*25*	94%	*35*	84%	*45*	97%
6	89%	*16*	91%	*26*	97%	*36*	96%	*46*	93%
7	86%	*17*	97%	*27*	86%	*37*	95%	*47*	92%
8	92%	*18*	88%	*28*	95%	*38*	94%	*48*	96%
9	86%	*19*	86%	*29*	89%	*39*	93%	*49*	96%
10	84%	*20*	84%	*30*	89%	*40*	89%	*50*	93%

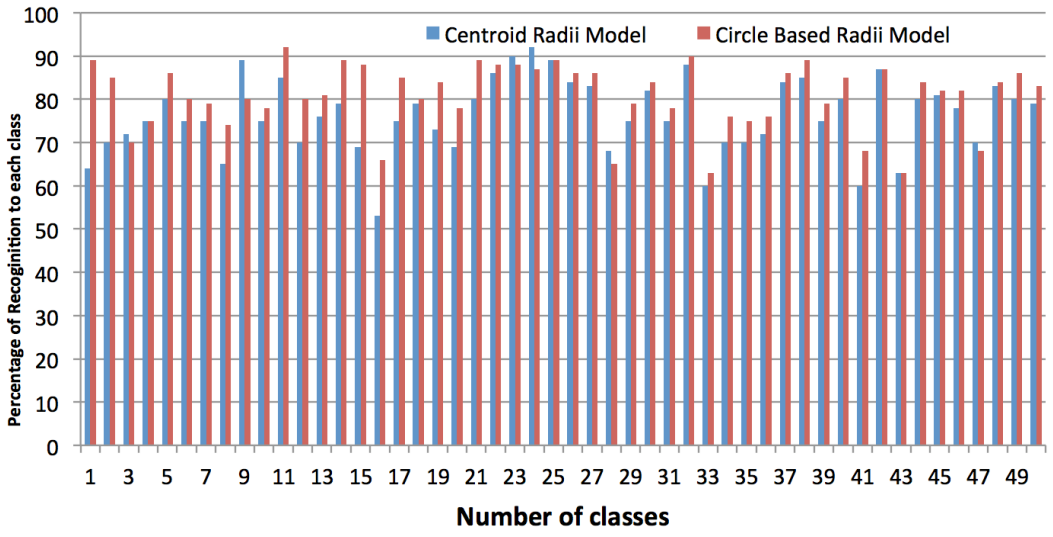


Figure 12. Performance comparison graph of the proposed circle-based radii model with the centroid radii model

Shown in Figure 13 is the 3D view-confusion matrix of the proposed circle-based radii model. Shown in Figure 14 is the 3D view of positive classification rate of the circle-based radii model. Shown in Figure 15 is the accuracy of the proposed circle-based Radii model.

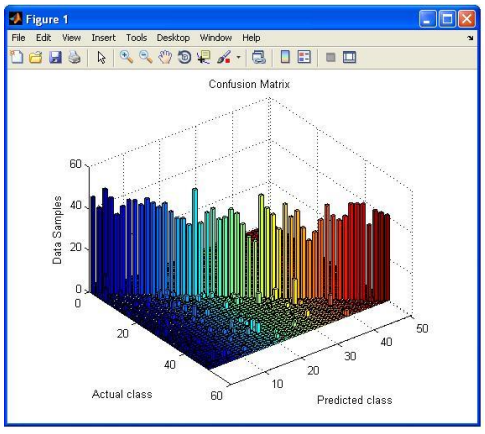


Figure 13. 3D view-confusion matrix of circle based radii model

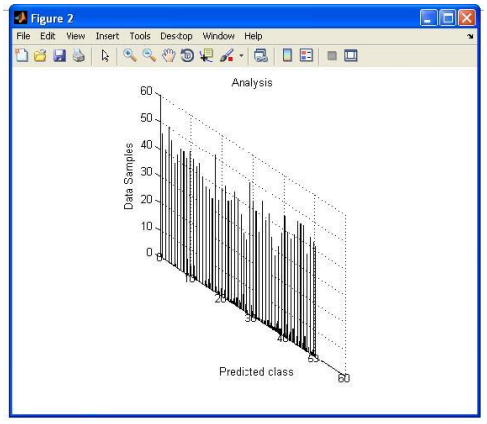


Figure 14. 3D view of positive classification rate of circle-based radii model

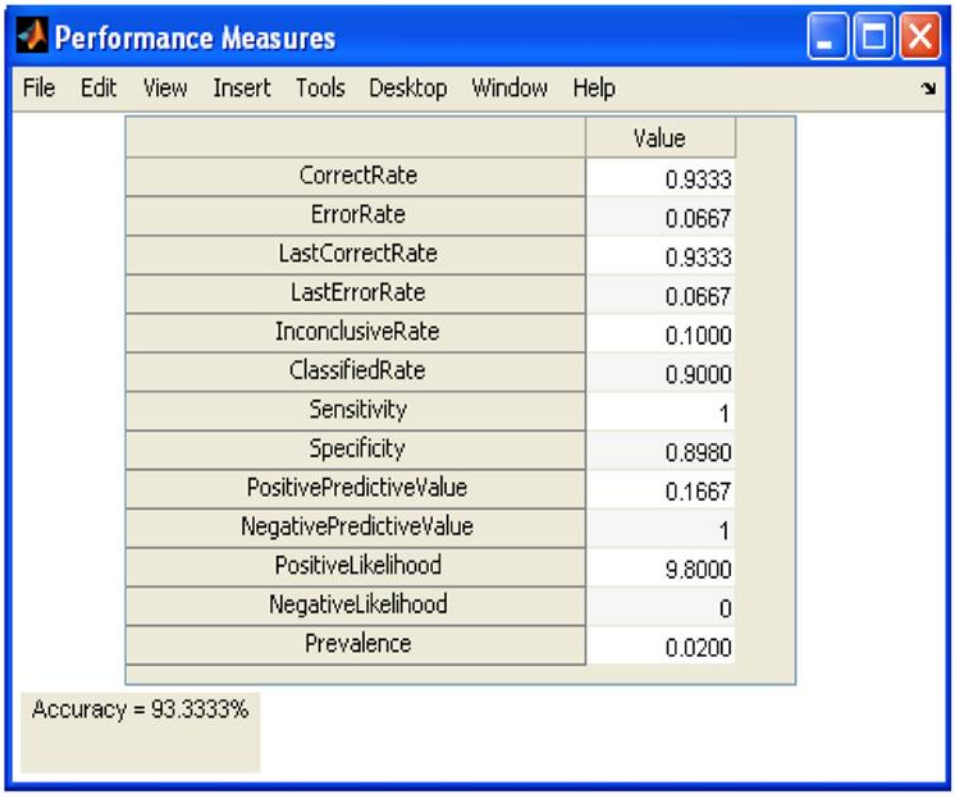


Figure 15. Accuracy of proposed circle-based radii model

Conclusion

One of the most important low-level features in content-based image retrieval is the shape. In this study a new shape descriptor, the circle-based radii model, was proposed. It is based on the center point and border of a circle centered inside the contour of a leaf. It is a successful feature extraction technique for the plant leaf classification system.

In the conventional method, the centroid radii model, the lengths of the radii from the centroid to the boundary are used to represent the shape, and the angular interval between radii is fixed. This conventional radii model generates a vector that is the normalized length of radii for shape representation. The vector depends on the order of the radius.

The proposed shape descriptor differs from the conventional centroid radii model in such a way that the distance will be calculated between the contour of

the circle and contour of leaf image, instead of calculating distance between the center point and border of the leaf. In this method, 44 features are extracted from every leaf in specified regions and compared with standard database features of the trained leaf images by using the SVM classifier.

By using the proposed method, the plant has been identified successfully in a large number of classes. The proposed method upgrades the set of feature values. The experimental results indicate that the proposed method shows significant improvement in terms of the increased number of features, and enhances the feature value. Accuracy of proposed circle-based radii model for the shape descriptor is 93.33% (the centroid radii model achieves 85.92% accuracy), indicating that the circle-based radii model is more suitable for a plant leaf classification system given its high retrieval performance.

References

Bong, M. F., Sulong, G., Kumoi, R. & Rahim, M. S. M. (2015). Classify the plant species based on lobes, sinuses and margin. *Jurnal Teknologi, 75*(2), 27–31. doi: 10.11113/jt.v75.4973

Bresenham, J. E. (1977). A linear algorithm for incremental digital display of circular arcs. *Communications of the ACM, 20*(2), 100–106. doi: 10.1145/359423.359432

Chaki, J. & Parekh, R. (2011). Plant leaf recognition using shape based features and network classifiers. *International Journal of Advanced Computer Science and Applications, 2*(10). doi: 10.14569/ijacsa.2011.021007

Deokar, S. R., Zope, P. H., & Suralkar, S. R. (2013). Leaf recognition using feature point extraction and artificial neural network. *International Journal of Engineering Research and Technology, 2*(1).

Fulsoundar, K., Kadlag, T., Bhadale, S., Bharvirkar, P. & Godse, S. P. (2014). Detection and classification of plant leaf diseases. *International Journal of Engineering Research and General Science, 2*(6), 868–874.

Intelligent Computing Laboratory, Chinese Academy of Science. (n.d.) *Intelligent Computing Laboratory (ICL) data set* [data set].

Lee, K. B. & Hong, K. S. (2013). An implementation of leaf recognition system using leaf vein and shape. *International Journal of Bio-Science and Bio-Technology, 5*(2), 57–66.

Ray, B. K. (2006). An alternative approach to circle drawing. *Journal of the Indian Institute of Science, 86*, 617–623.

Tan, K. L., Ooi, B. C. & Thiang, L. F. (2003). Retrieving similar shapes effectively and efficiently. *Multimedia Tools and Applications, 19*(2), 111–134. doi: 10.1023/a:1022142527536

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 481-497.
doi: 10.22237/jmasm/1493598480

ISSN 1538 − 9472

Distribution Fits for Various Parameters in the Florida Public Hurricane Loss Model

Victoria Oxenyuk
JM Family Enterprises, Inc
Deerfield Beach, FL

Sneh Gulati
Florida International University
Miami, FL

B M Golam Kibria
Florida International University
Miami, FL

Shahid Hamid
Florida International University
Miami, FL

The purpose of this study is to re-analyze the atmospheric science component of the Florida Public Hurricane Loss Model v. 5.0, in order to investigate if the distributional fits used for the model parameters could be improved upon. We consider alternate fits for annual hurricane occurrence, radius of maximum winds and the pressure profile parameter.

Keywords: Gamma distribution, goodness-of-fit, hurricanes model, normal distribution, Poisson distribution, Weibull

Introduction

Hurricanes are one of the greatest natural hazards; relatively rare in occurrence but capable of causing colossal economic losses. In 1992, "when Hurricane Andrew struck Florida it caused over $30 billion in direct economic losses" (Lokupitiya, Borgman, & Anderson-Sprecher, 2005, p. 4394). Hurricane modeling has become a widely used tool for assessing risks associated with windstorm catastrophes. Since the groundbreaking studies of Russell (1968, 1971) and Tryggvason, Davenport, and Surry (1976), the modeling methods have improved significantly due to increased computing capabilities, new advanced physical and statistical models, and vast growth in quantity and quality of available data. Several private models for simulating hurricane loss have been

Victoria Oxenyuk is a Reporting Analyst. Sneh Gulati is a Professor in the Department of Mathematics and Statistics. Email them at gulati@fiu.edu. B M Golam Kibria is a Professor in the Department of Mathematics and Statistics. Email them at kibriag@fiu.edu. Shahid Hamid is a Professor in the Department of Finance. Email them at hamids@fiu.edu.

developed in the recent years for use in the State of Florida, but such models typically are commercial and are not available to the research community and public. The Florida Public Hurricane Loss Model (FPHLM) is a notable exception.

The FPHLM is an open public hurricane loss evaluation model, which was developed jointly by specialists in the fields of meteorology, engineering, computer science, finance, and statistics from the Florida State University system (SUS), National Oceanic and Atmospheric Administration (NOAA) Hurricane Research Division, and the University of Miami. This model was created "for the purpose of probabilistic assessment of risk to insured residential property associated with wind damage from hurricanes" (Hamid et al., 2005, p. 552).

FPHLM consists of three main components: first, the atmospheric science component which models the track and intensity of hurricanes that threaten Florida; second, the engineering component which models vulnerability of insured property; and third, the actuarial science component which models the insured loss. In order to be used for rate making purposes in the State of Florida, a model has to the rigorous statistical standards set by the Florida Commission for Hurricane Loss Projection Methodology (FCHLPM.) The purpose of this study it to re-analyze some of the components of the atmospheric component of the FPHLM v 5.0 model certified by the commission in 2011.

The atmospheric science component simulates thousands of storms, their wind speeds, and their decay once on land based on historical hurricane statistics, thus defining probabilistic wind risk for all residential zip codes in Florida. The wind risk information is then passed on to the engineering and actuarial science components to assess damage and annual insured loss. Each component is developed independently and delivered as a one-way input to the next component in line until the end result is achieved. We now look at the atmospheric science component in details.

The first step in modeling annual wind risk for a zip code is the determination of a model for the annual hurricane occurrence (AHO). FPHLM uses a non-parametric method to estimate annual hurricane occurrence, in that we sample from historical records to determine the number of hurricanes in a given year. The research question was if a parametric distribution could be used to estimate AHO instead. The two alternative distributions were the Poisson distribution that assumes homogenous hurricane frequencies (the mean number of hurricanes in any two years is the same) or the Negative Binomial distribution that assumes a non-homogenous annual occurrence rate.

In addition to investigating fits for AHO, it was also decided to reanalyze two other important storm parameters, radius of maximum winds, R_{max}, and the

pressure profile parameter, Holland *B*. These two variables are important for estimating loss. Greater values of the radius of maximum winds imply greater losses and, similarly, lower values of central pressure mean a more intense hurricane and therefore higher losses.

The sensitivity and uncertainty analysis shows that loss costs are fairly sensitive to Holland *B* and R_{max} regardless of hurricane category. FPHLM has historically used the Gamma distribution to fit R_{max}. The question arose, however, if there were other distributions that might provide better fits for R_{max}.

Holland *B* is an additional parameter defining the pressure field and maximum wind speeds in a hurricane. It was introduced by Holland (1980) and has been used in many hurricane threat studies since. FHPLM shows that the Holland *B* parameter is inversely correlated with both the size and latitude of the hurricane. Here we investigate alternate models for Holland *B* and see if they explain more of the variability in Holland *B* as compared to the present model.

As specified by the FCHLPM, analysis of annual hurricane occurrence and radius of maximum winds (for PHLM v 5.0) is based on the data obtained from historical record for the Atlantic tropical cyclone basin (known as HURDAT) for the period from 1901 till 2010. Earlier data is available but not used due to lack of population centers and uncertainties about meteorological measurements before the start of 20th century. The model for the Holland *B* pressure profile parameter is developed based on a subset of the data published by Willoughby and Rahn (2004) and obtained by NOAA and U.S. Air Force Reserve aircraft between 1977 and 2000.

To find the best fitting distribution, a preliminary analysis of the data was conducted through the use of EasyFit software which allows us to easily fit a large number of distributions to the data. Estimated parameters of the best fitting distributions were then found using the maximum likelihood estimator (MLE) method. In order to determine how well the selected distributions fit the data, they were tested for goodness-of-fit using Kolmogorov-Smirnov, Anderson-Darling, and Chi-Squared tests. Along with the goodness-of-fit tests, the probability density function graphs, Q-Q, and P-P plots were also used to enable visual assessment of the goodness-of-fit and empirically compare several fitted models. In order to determine the model for the estimation of Holland *B*, multiple regression analysis was performed using the PROC REG procedure in SAS.

Annual Hurricane Occurrence

The first step in the study of hurricanes and their impacts is to determine the frequency with which they occur. Annual Hurricane Occurrence (AHO) rate estimates "the frequency of hurricanes occurring in a series of years based on an associated hurricane occurrence probability distribution, which is obtained through statistical analysis and calculation on the basis of historical hurricane records" (Chen et al., 2004, p. 6). In the recent years, substantial research in the area of modeling the occurrence of hurricanes has been done by Chen et al. (2003, 2004), Gray, Landsea, Mielke, and Berry (1992), Elsner and Schmertmann (1993), and Elsner and Jagger (2004). The basic principle of these papers was to develop the statistical models from the available historical data in order to estimate AHO. Based on the obtained probability distributions, the number of hurricanes per year in the future is produced for a desired number of years.

The Poisson and the Negative Binomial distributions are often used by modeling agencies to model AHO. The rate of occurrence of a stochastic process is typically described by the use of the Poisson distribution. However, Poisson distributions assume the mean number of storms in any two non-overlapping time intervals of the same length to be equal. To allow those means to be unequal will lead to the modeling of the annual occurrence by the Negative Binomial distribution. General guiding principles as to the adequacy of the two distributions have been discussed (Thom, 1966), but one cannot accurately determine which model is appropriate until necessary tests are conducted. In this section we determine whether the Poisson or the Negative Binomial is adequate in describing the distribution of the annual hurricane occurrence.

For the assessment of the AHO distribution to be conducted, a suitable data set has to be obtained. Annual counts of tropical storms and hurricanes in the Atlantic Ocean are obtained from the HURDAT (National Oceanic and Atmospheric Administration's Hurricane Research Division, 2012) database, which is maintained by the National Hurricane Center in Miami, Florida and the National Climatic Data Center in Asheville, North Carolina. This historical record for the Atlantic tropical cyclone basin contains positions and intensities of tropical storms and hurricanes recorded every six hours from 1851 onwards. However, as specified by the commission, we use data starting from 1901 for our research due to the unreliability of 19th century data. At the time as this research was conducted, the FPHLM was based on the period 1901-2010, thus all our analysis is conducted on the HURDAT data from 1901-2010. In its analysis of the hurricane counts, FPHLM does not count all hurricanes in the Atlantic. Instead, it counts only the

storms in a "threat area" (Figure 1) – within 1000 km of a location (26.0 N, 82.0 W) – in order to focus on storms capable of affecting residential property in Florida.

In order to obtain the number of hurricanes in each year from 1901 to 2010, FPHLM looks at each hurricane and its six hourly positions recorded by HURDAT. The first time a hurricane entered the threat area during its track was counted as an occurrence. Subsequent entries by the same storm were not counted, so that any hurricanes could only be counted once. The annual number of hurricanes in any given year range between 0 and 5 with mean 1.1091 and standard deviation 1.1704, as seen in the summary statistics for AHO in Table 1.

Each storm is considered as a point event in time, occurring independently. If λ is a measure of the historically based number of events per year, then $P(X = x \mid \lambda)$ defines the probability of having x events per year, which is given by the Poisson probability distribution function (PDF)

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

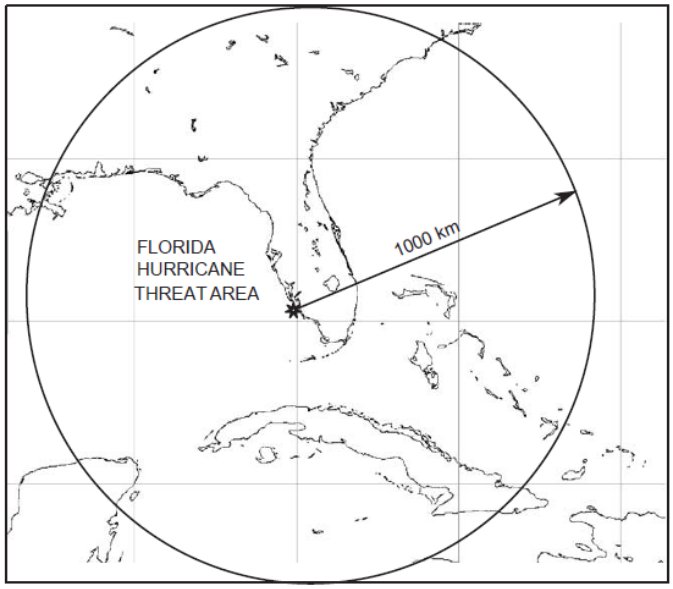


Figure 1. Florida hurricane threat area

Table 1. Descriptive statistics of annual occurrence rate

Sample Size (*N*)	110	Min	0
Mean	1.1091	Median	1
Variance	1.3699	Max	5
Std Deviation	1.1704	Range	5

The parameter λ of the Poisson distribution can be estimated from data by the maximum likelihood estimator

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} x_i}{N}$$

where x_i is the number of events in a given year and N is the total number of years.

However, if it is assumed that the number of events X has a Negative Binomial distribution, then the corresponding pdf for the distribution is given by

$$\mathrm{P}(x) = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}\left(\frac{k}{m+k}\right)^{k}\left(\frac{m}{m+k}\right)^{x}$$

where Γ is the gamma function and m and k are parameters of the distribution. The MLEs of the parameters m and k can be obtained as

$$\hat{m} = \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{and} \quad \hat{k} = \frac{\hat{m}^2}{s^2 - \hat{m}}$$

where s^2 is the sample variance.

The parameters of both the Poisson and Negative Binomial distributions were estimated using annual number of hurricanes dataset and results are presented in Table 2.

Table 2. Estimated parameters of the distribution for AHO data

Distribution	**Parameter Values**
Poisson	$\lambda = 1.1091$
Negative Binomial	$n = 4$, $p = 0.8096$

Note: The parameters of the negative binomial distribution are $n = k + m$ and $p = k/(m + k)$

Table 3. Goodness-of-fit tests for AHO data

	Chi-Squared			Kolmogov-Smrinov		Anderson-Darling	
Distribution	**Statistic**	***p*-value**	**Rank**	**Statistic**	**Rank**	**Statistic**	**Rank**
Poisson	1.71979	0.88640	1	0.32986	1	16.465	1
Neg. Binomial	2.83815	0.58527	2	0.42963	2	28.094	2

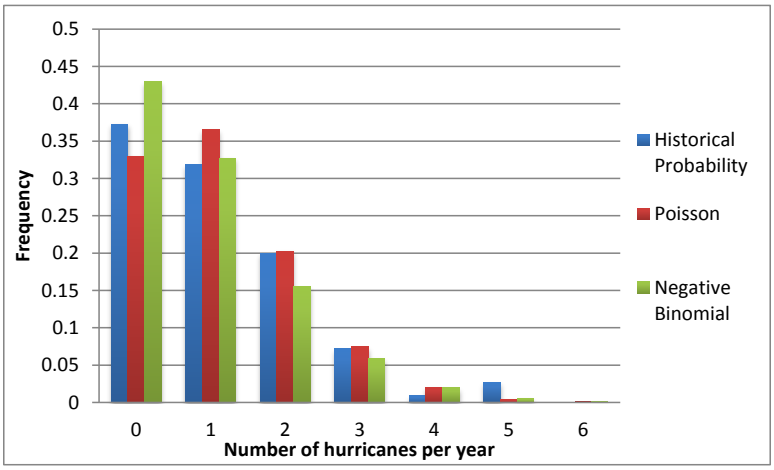


Figure 2. Comparison of simulated vs. historical occurrences

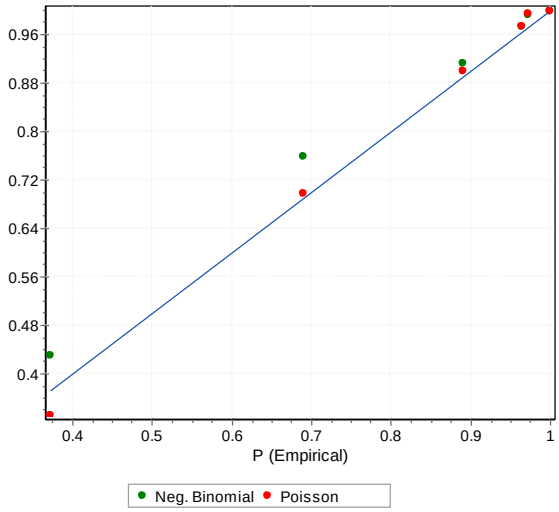


Figure 3. P-P plot

Once distributions were fitted, it was decided to conduct goodness-of-fit tests to see which distribution provided a better fit. The tests considered were the Kolmogorov-Smirnov test, the chi-square test, and the Anderson-Darling test. The results are given in Table 3. It is clear the Poisson distribution provides a better fit for AHO using the threat area.

The distribution graphs were examined to provide a visual assessment and an empirical comparison of the goodness-of-fit. Indicated in Figure 2 are the occurrence rates of historical and modeled hurricane data. A P-P plot of the fitted distributions is presented in Figure 3. It is not clear from Figure 2 which distribution provides a better fit, but Figure 3 does make it clear that the Poisson distribution is a better fit in keeping with the goodness-of-fit tests.

It was concluded the best fitting distribution for the annual hurricane occurrence for the Florida threat area, based on the results of goodness-of-fit tests and the P-P plot, is a Poisson distribution with parameter $\lambda = 1.1091$.

Radius of Maximum Winds

Consider the wind field model for the FPHLM; specifically, consider the radius of maximum winds at landfall, the distance between the center of a cyclone and its band of strongest winds. Meteorologists at FPHLM developed an R_{max} model using a landfall R_{max} database from Ho, Su, Hanevich, Smith, and Richards (1987) and supplemented by NOAA HRD research flight data and NOAA-HRD H*Wind analyses (Powell et al., 2005). The current database includes 112 measurements of radius of maximum wind, central pressure, and location at landfall for storms from 1901 till 2010.

Values of R_{max}, measured in statute miles, range between 5.75 and 52.9 with mean 25.65 and standard deviation 11.2 as seen in Table 4.

The histogram of the data is depicted in Figure 4 and shows that the R_{max} data is right-skewed. A preliminary analysis of the R_{max} landfall database was conducted using the Easyfit software. As initial models, we considered right-skewed distributions with a maximum of 2 parameters (extra parameters would have made the use for the wind field model over-complicated and not practical). Moreover, it was desirable to avoid the situations where distributions with more parameters may well fit the data better because of a lot more flexibility in shape, but then the apparent improvement would be spurious due to over-fitting.

Table 4. Descriptive statistics of radius of maximum winds

Sample size	112	Min	5.75
Mean	25.649	Median	24.725
Variance	125.31	Max	52.9
Std. deviation	11.194	Range	47.15

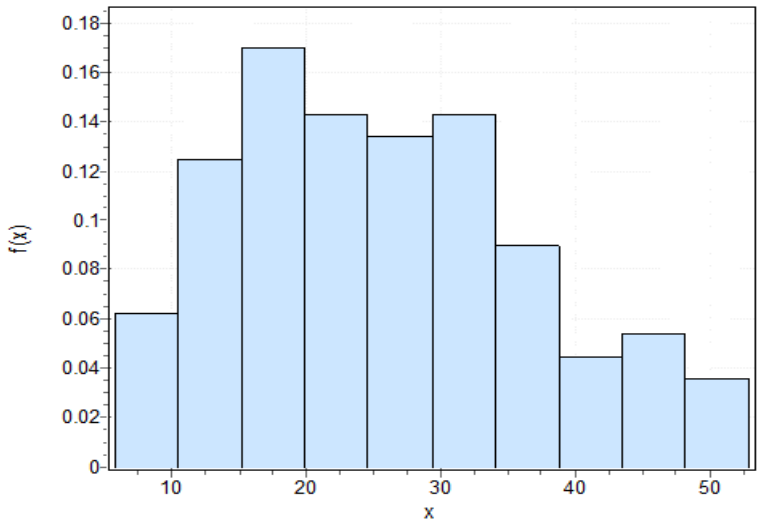


Figure 4. Probability density function radius of maximum winds

Five distributions that were found to be a good fit for modeling R_{max} based on the above criteria were Gamma, Lognormal, Rayleigh, Weibull, and Inverse Gaussian. Gamma and Lognormal are the distributions that were considered in the FPHLM and Gamma was chosen as the best fit. Parameters of selected distributions were obtained using MLEs and results are presented in the Table 5.

Once again, they were tested for goodness-of-fit in order to determine how well the selected distributions fit the $R_{\max}$ data. Due to the continuous nature of the data and the low power of the chi-squared test, the Anderson-Darling and the Kolmogorov-Smirnov tests were employed. They were chosen because they are general, apply to all continuous distributions, and have high power. The results are presented in Table 6.

The distributions are ranked according to the *p*-value of the test, with higher *p*-values indicating a better fit. Regardless of the test being used, both the Lognormal and Inverse Gaussian distributions show a poor fit for $R_{\max}$ data with

p-values below 0.5 for the K-S test. It was concluded that Lognormal and Inverse Gaussian distributions are not good fits and exclude them from further consideration.

The three distributions for be considered further are Weibull, Rayleigh and Gamma. Gamma distribution is used to fit the radius of maximum winds in the Florida Public Hurricane Loss Evaluation Model, however, notice both the Weibull and Rayleigh perform better than the Gamma distribution according to the tests.

In order to finalize the model, a visual inspection of the data set was conducted starting with the Probability Density Function Graph for the data. The graph displays the theoretical PDFs of the fitted distributions and the histogram of the R_{max} data (Figures 5 and 6). Because the histogram depends on how the data is sorted into bins, two histograms are displayed with the Rmax values binned in 10 and 15 intervals for comparative analysis. All three distributions are plotted on the same graphs. Displaying several distributions at the same time will allow us to visually compare the models and determine how they differ.

Although it is hard to make a decision about better fit based on these graphs as they require the arbitrary grouping of the data, Weibull and Rayleigh distributions do appear to fit the data better.

Table 5. Estimated distribution parameters for R_{max} data

Distributions	Parameters
Gamma	α = 5.250, β = 4.886
Lognormal	δ = 0.492, μ = 3.136
Weibull	α = 2.474, β = 28.666
Raleigh	δ = 17.293, γ = 3.879
Inverse Gamma	λ = 134.66, μ = 25.650

Table 6. Goodness-of-fit tests for R_{max} data

	Kolmogov-Smrinov			Anderson-Darling	
Distributions	Statistic	*p*-value	Rank	Statistic	Rank
Weibull	0.0494	0.9349	1	0.3226	1
Rayleigh	0.0561	0.8530	2	0.3006	2
Gamma	0.0703	0.6124	3	0.5349	3
Lognormal	0.0904	0.3015	4	1.0419	4
Inverse Gaussian	0.0953	0.2450	5	1.8773	5

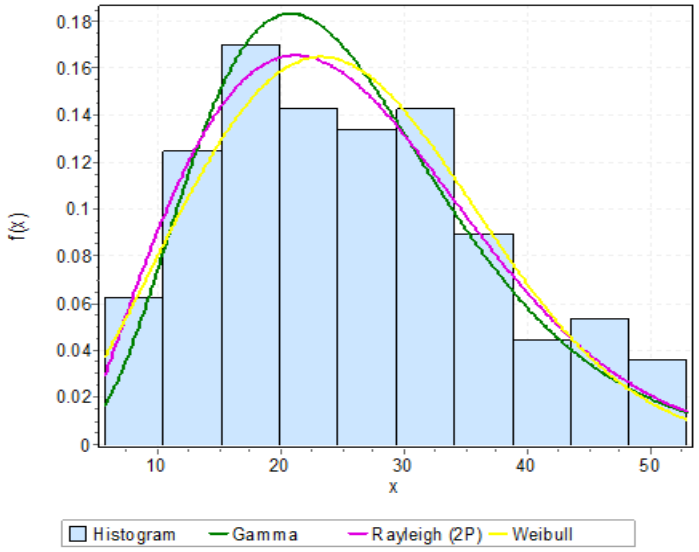


Figure 5. PDF graph with R_{max} values binned in 10 intervals

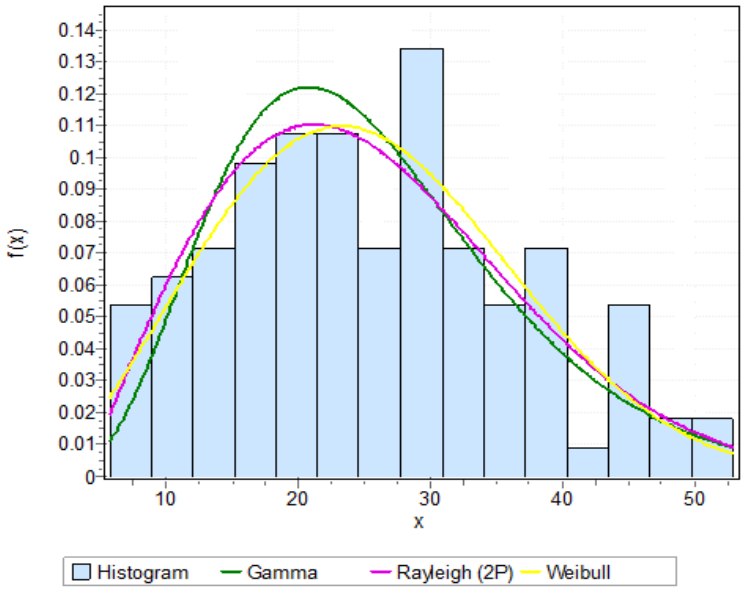


Figure 6. PDF graph with R_{max} values binned in 15 intervals

To avoid grouping of the data, consider the Q-Q plot (Figure 7). Although all three distributions appear to be good fits based on the Q-Q plot, it appears that the Gamma and Rayleigh distributions have points further away from the straight line as values of R_{max} get larger. This is consistent with the results of the Kolmogorov-Smirnov test. Based on the results of the goodness-of-fit test, the PDF graph, and the Q-Q plot, it was concluded the Weibull distribution with parameters $\alpha = 2.4736$ and $\beta = 28.666$ is the best fit for the Radius of maximum winds.

Although it was shown that the Weibull distribution provided a better fit for R_{max} based on the data set, the Gamma distribution was used for modeling the radius of maximum winds in the FPHLM. The analysis shows the Gamma distribution as a possible fit for the radius of maximum winds, although perhaps not the best fit. Both the Gamma and Weibull distributions are commonly encountered in reliability analysis and it is often difficult to choose between the two. Hence, it should be stressed the Gamma distribution was not rejected as a possible fit for R_{max}. Instead, it was concluded the Weibull might be a better fit.

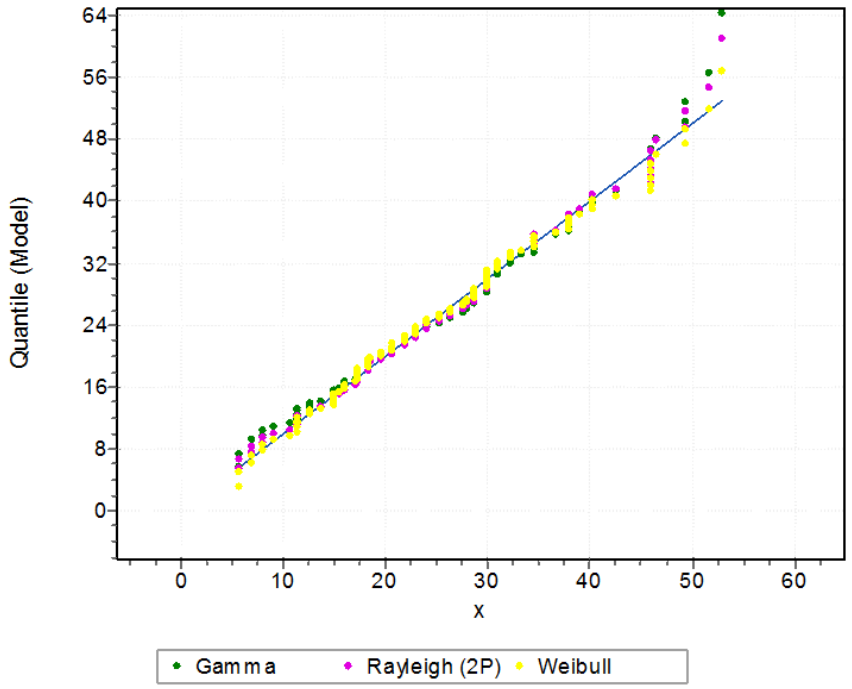


Figure 7. Q-Q plot

Holland *B*

Another important parameter of the wind field model is the Holland *B* parameter. Holland *B* is an additional parameter defining the pressure field and maximum wind speeds in a hurricane. It was introduced by Holland in 1980 and has since been used in hurricane threat studies by many researchers including Powell et al. (2005), James and Mason (2005), Emanuel, Ravela, Vivant, and Risi (2006), Lee and Rosowsky (2007), Hall and Jewson (2008), Vickery and Wadhera (2008), and Vickery, Masters, Powell, and Wadhera (2009), among others. The relation between the pressure of a hurricane, p(*r*), and the Holland *B* parameter is given as follows:

$$\mathrm{p}(r) = p_c + \Delta p \mathrm{e}^{-\left(\frac{R_{\max}}{r}\right)^B}$$

where *r* is the distance from the center of the storm, p_c is the pressure at the center of the storm, Δp is the difference between central minimum sea level pressure (p_c) and the outer peripheral pressure (1013 mb), and $R_{\max}$ is the radius of maximum winds. Thus Holland *B* allows for the distinction in the maximum wind speeds observed in hurricanes for a given Δp (all else being equal). With the introduction of the *B* parameter, the maximum wind speeds in the simulated hurricane are proportional to $\sqrt{B\Delta p}$ compared to $\sqrt{\Delta p}$ otherwise.

In meteorological literature, Holland *B* is often modeled as a linear function of the location of the storm, the radius of maximum winds, and the central pressure difference or deficit Δp. FPHLM uses a similar regression fit for Holland *B* based on a filtered subset of the data published by Willoughby and Rahn (2004). The data consist of winds and geo-potential heights obtained by the NOAA and U.S. Air Force Reserve aircraft between 1977-2000, supplemented with Δp, the pressure deficit, and $R_{\max}$ values. FPHLM retains 116 profiles filtered as follows:

1) by Height of flight-level pressure surface ≤ 700,
2) Longitude between 70 and 95 degrees west,
3) Storm relative flight level Vmax > 33 m/s,
4) Latitude between 20 and 34 degrees North.

The final fitted model used by FPHLM is based on statistical analysis as well as validation using storm tracks and is

$$B = 1.74425 - 0.007915\,\mathrm{Lat} + 0.0000084\Delta p^2 + 0.005024 R_{\max} \quad (1)$$

This model explains about 15% of the variability in the four Holland *B*.

Most Holland *B* models have low R^2 values, and the model used by FPHLM does have higher R^2 values than most available models. It was decided to investigate if equation (1) could be further improved on in terms of a higher R^2 value by examining functions of Holland *B* other than liner functions or by the inclusion of other variables. Using the same data set as the one used by the FPHLM, we considered various fits for Holland *B* using latitude, longitude, Δp, and $R_{\max}$ as independent variables.

Matrix scatter plots indicated that using ln(*B*) as an dependent variable rather than *B* might yield better fits. However, a detailed stepwise regression analysis in SPSS did not yield a better fit when using ln(*B*) as a dependent variable. Stepwise regression indicates that the only variable significant in predicting either *B* or ln(*B*) is $R_{\max}$. Using *B* as a dependent variable yields an R^2 of 0.112 while using ln(*B*) as a dependent variable yields an R^2 of 0.122. Although it appears from the analysis there was no statistical need to use Δp or latitude in fitting Holland *B*, it is not recommended to make changes to the present fit for Holland *B* in the FPHLM; the analysis does not yield a better fit and the benefit of validating the fit using actual storms was not available.

Conclusion

The FPHLM is the only open public hurricane loss evaluation model available for the assessment of hazard to insured residential property related to damage from hurricanes in Florida. A numerical analysis of the atmospheric science component of the Florida Public Hurricane Loss Model was conducted to determine if it was possible to develop alternate models for the various hurricane parameters.

Based on the results of goodness-of-fit tests, histograms of historical and modeled occurrences, and P-P plots, it was concluded that the best fitting distribution for the annual hurricane occurrence is the Poisson distribution. The radius of maximum winds has a substantial impact on the area affected by hurricane and modeling of the $R_{\max}$ influences the likelihood of the location experiencing strong winds in cases of near misses. The Weibull was chosen as the best fit for the radius of maximum winds. The fit for Holland *B* being used by the FPHLM could not be improved. It was shown the models presented for Annual Hurricane Occurrence and $R_{\max}$ are better fits than the ones used by FPHLM, although it was not recommended the FPHLM change its modeling strategies. The

models considered by the FPHLM are consistent with models used in meteorological literature. However, this investigation might start a conversation in the meteorological community to search for alternate models for modeling hurricane parameters.

References

Chen, S.-C., Gulati, S., Hamid, S., Huang, X., Luo, L., Morisseau-Leroy, N.,… Zhang, C. (2003, August). *A three-tier system architecture design and development for hurricane occurrence simulation*. Paper presented at the International Conference on Information Technology: Research and Education. Newark, NJ.

Chen, S.-C., Gulati, S., Hamid, S., Huang, X., Luo, L., Morisseau-Leroy, N.,… Zhang, C. (2004). A Web-based distributed system for hurricane occurrence projection. *Software: Practice and Experience, 34*(6), 549-571. doi: 10.1002/spe.580

Elsner, J. B., & Jagger, T. H. (2004). A hierarchical Bayesian approach to seasonal hurricane modeling. *Journal of Climate, 17*(14), 2813-2827. doi: 10.1175/1520-0442(2004)017<2813:AHBATS>2.0.CO;2

Elsner, J. B., & Schmertmann, S. C. (1993). Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Weather and Forecasting, 8*(3), 345-351. doi: 10.1175/1520-0434(1993)008<0345:IERSPO>2.0.CO;2

Emanuel, K. A., Ravela, S., Vivant, E., & Risi, C. (2006). A statistical-deterministic approach to hurricane risk assessment. *Bulletin of the American Meteorological Society, 87*(3), 299-314. doi: 10.1175/BAMS-87-3-299

Gray, W. M., Landsea, C. W., Mielke, P. W., Jr., & Berry, K. J. (1992). Predicting Atlantic seasonal hurricane activity 6-11 months in advance. *Weather and Forecasting, 7*(3), 440-455. doi: 10.1175/1520-0434(1992)007<0440:PASHAM>2.0.CO;2

Hall, T., & Jewson, S. (2008). Comparison of local and basin-wide methods for risk assessment of tropical cyclone landfall. *Journal of Applied Meteorology and Climatology, 47*(2), 361-367. doi: 10.1175/2007jamc1720.1

Hamid, S., Kibria, B. M. G., Gulati, S., Powell, M., Annane, B., Cocke, S.,… Chen, S.-C. (2005). Predicting losses of residential structures in the state of

Florida by the public hurricane loss evaluation models. *Statistical Methodology, 7*(5), 552-573. doi: 10.1016/j.stamet.2010.02.004

Ho, F. P., Su, J. C., Hanevich, K. L., Smith, R. J., & Richards, F. P. (1987). *Hurricane climatology for the Atlantic and Gulf Coasts of the United States*. (NOAA Technical Report NWS 38). Silver Spring, MD: U. S. Department of Commerce. Retrieved from https://coast.noaa.gov/hes/images/pdf/ATL_GULF_HURR_CLIMATOLOGY.pdf

Holland, G. J. (1980). An analytic model of the wind and pressure profiles in hurricanes. *Monthly Weather Review, 108*(8), 1212-1218. doi: 10.1175/1520-0493(1980)108<1212:AAMOTW>2.0.CO;2

James, M. K., & Mason, L. B. (2005). Synthetic tropical cyclone database. *Journal of Waterway, Port, Coastal, and Ocean Engineering, 131*(4), 181-192. doi: 10.1061/(asce)0733-950x(2005)131:4(181)

Lee, K. H., & Rosowsky, D. V. (2007). Synthetic hurricane wind speed records: Development of a database for hazard analyses and risk studies. *National Hazards Review, 8*(2), 23-34. doi: 10.1061/(asce)1527-6988(2007)8:2(23)

Lokupitiya, R., Borgman, L., & Anderson-Sprecher, R. (2005). Simulation of storm occurrences using simulated annealing. *Journal of Climate, 18*(21), 4394-4403. doi: 10.1175/jcli3546.1

National Oceanic and Atmospheric Administration's Hurricane Research Division. (2012). *Easy to read HURDAT 2012* [Data file]. Retrieved from http://www.aoml.noaa.gov/hrd/hurdat/easyread-2012.html

Powell, M. D., & Houston, S. H. (1996). Hurricane Andrew's landfall in south Florida. Part II: Surface wind fields and potential real-time applications. *Weather and Forecasting, 11*(3), 329-349. doi: 10.1175/1520-0434(1996)011<0329:HALISF>2.0.CO;2

Powell, M. D., & Houston, S. H. (1998). Surface wind fields of 1995 Hurricanes Erin, Opal, Luis, Marilyn, and Roxanne at landfall. *Monthly Weather Review, 126*(5), 1259-1273. doi: 10.1175/1520-0493(1998)126<1259:SWFOHE>2.0.CO;2

Powell, M. D., Soukup, G., Cocke, S., Gulati, S., Morisseau-Leroy, N., Hamid, S.,… Axe, L. (2005). State of Florida hurricane loss projection model: Atmospheric science component. *Journal of Wind Engineering and Industrial Aerodynamics, 93*(8), 651-674. doi: 10.1016/j.jweia.2005.05.008

Russell, L. R. (1968). *Probability distribution for Texas gulf coast hurricane effects of engineering interest* (Doctoral dissertation). Stanford University, Stanford, CA.

Russell, L. R. (1971). Probability distributions for hurricane effects. *Journal of Waterways, Harbors & Coastal Engineering Division, 97*(WW1), 139-154.

Thom, H. C. S. (1966). *Some methods of climatological analysis*. World Meteorological Organization Technical Note No. 81.

Tryggvason, V. J., Davenport, A. G., & Surry, D. (1976). Predicting wind-induced response in hurricane zones. *Journal of the Structural Division, 102*(ST12), 2333-2350.

Vickery, P. J., Masters, F., Powell, M., & Wadhera, D. (2009). Hurricane hazard modeling: The past, present, and future. *Journal of Wind Engineering and Industrial Aerodynamics, 97*(7-8), 392-405. doi: 10.1016/j.jweia.2009.05.005

Vickery, P. J., & Wadhera D. (2008). Statistical models of Holland pressure profile parameter and radius to maximum winds of hurricanes from flight-level pressure and H*wind data. *Journal of Applied Meteorology and Climatology, 47*(10), 2497-2517. doi: 10.1175/2008jamc1837.1

Willoughby, H. E., & Rahn, M. E. (2004). Parametric representation of the primary hurricane vortex. Part I: Observations and evaluation of the Holland (1980) model. *Monthly Weather Review, 132*(12), 3033-3048. doi: 10.1175/mwr2831.1

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 498-517.
doi: 10.22237/jmasm/1493598540

ISSN 1538 − 9472

Multivariate Multilevel Modeling of Age Related Diseases

Kapuruge N. O. Ranathunga
University of Colombo
Colombo, Sri Lanka

Roshini Sooriyarachchi
University of Colombo
Colombo, Sri Lanka

The emerging role of modeling multivariate multilevel data in the context of analyzing the risk factors are examined for the severity of cardiovascular disease diabetes, and chronic respiratory conditions. The modeling phase results leads to some important interaction terms between blood glucose, blood pressure, obesity, smoking and alcohol to the mortality rates.

Keywords: multivariate multilevel model, probit regression, cardiovascular disease and diabetes, chronic respiratory conditions, markov chain Monte Carlo

Introduction

Aging increases susceptibility to age-associated diseases and some of these diseases may increase mortality among adults worldwide. The focus of this study is on cardiovascular disease and diabetes (CDD) and chronic respiratory conditions (CRC). These are life-threatening diseases with increasing incidence. Also, there is a geographical effect of the mortality rates of these diseases (World Health Organization, 2005). Therefore, countries are grouped into continents geographically, but vary across continents. This establishes the need of multilevel hierarchical analysis

Because the existence of a high correlation between these variables, and the presence of some common risk factors to these related diseases, a multivariate multilevel concept was used to identify the joint effects of some risk factors on these two diseases to analyze data more appropriately. A multivariate multilevel model can be considered as a collection of multiple dependent variables in a hierarchical nature. When the effect of a set of explanatory variables on a set of dependent variables shows a considerable difference then it can be handled only by means of a multivariate analysis (Snijders & Bosker, 2012).

Kapuruge N. O. Ranathunga is a former Assistant Lecturer in the Department of Statistics. Email her at nishikaoshadini@gmail.com.

Table 1. Description of the data and its abbreviations

Variable Name	Identifier	Category	Code
Cardiovascular Diseases and Diabetes (per 100,000 population)	CDD	<220	1
		220-370	2
		>370	3
Chronic Respiratory Conditions (per 100,000 population)	CRC	<20	1
		20-50	2
		>50	3
Population using improved drinking water sources (%)-2011[a]	Water	<88	1
		88-98	2
		>98	3
Population using improved sanitation (%)-2011[a]	Sanitation	<40	1
		40-80	2
		>80	3
Population using solid fuels (%)-2011[a]	Solid_Fuel	<20	1
		20-70	2
		>70	3
Prevalence of raised fasting blood Glucose among adults aged ≥ 25 years (%)-2008[a]	B_Glucose	<7.5	1
		7.5-11.5	2
		>11.5	3
Prevalence of raised blood pressure among adults aged ≥ 25 years (%)-2008[a]	B_Pressure	<25	1
		25-35	2
		>35	3
Adults aged ≥ 20 years who are obese (%)-2008[a]	Obese	<13	1
		13-24	2
		>24	3
Alcohol consumption among adults aged ≥ 15 years (litres of pure alcohol per person per year)-2008[a]	Alcohol	<4	1
		4-10	2
		10-16	3
		>16	4
Prevalence of smoking any tobacco product among adults aged ≥ 15 years (%)-2009[a]	Smoking	<12	1
		12-24	2
		24-36	3
		>36	4

Note: *a*) country level (1st level) variables

Considered here is a multivariate multilevel analysis approach by using Bayesian methods. Data for this study were obtained from the World Health Organization (2013). The dataset consists of worldwide mortality rates among

adults aged 30-70 years. Due to the incompleteness of the records, a multiple imputation (MI) was conducted to variables Smoking, Water and Sanitation prior to fitting the models (Sterne et al., 2009). The MI procedure requires the variables to be imputed to be normally distributed or categorical. Water and Sanitation did not follow a normal distribution, and were categorized to perform the MI (Table 1), and are considered ordinal categorical variables for the modeling.

Given in Table 1 are the variables and their respective categories with abbreviations. The continuous data were discretized in to ⅓ splits based on percentiles to obtain respective categories. Alcohol and smoking were categorized into ¼ splits to obtain more explicate categories due to the expansion of the data.

Methodology

Univariate analysis using Zhang and Boos test

Before carrying out the modeling it is essential to determine the nature of the strength of the relationships between explanatory variables and response variables. However due to the natural hierarchy of the observations, Zhang & Boos (1997) developed the Generalized Cochran Mantel Haenszel (GCMH) test. There are three different types of test statistics proposed by Zhang and Boos. These are T_{EL}, T_P and T_U. From these, T_P is preferable to T_U and T_{EL} (Jayawardana and Sooriyarachchi, 2014). Simulation studies showed it maintains error values even for a small number of strata (Zhang and Boos, 1997).

Structure of the Multivariate Multilevel Probit Regression Model

Although the logit link is the most common, the multivariate model for binary responses was developed for the probit link in MLwiN 2.10 (Rasbash et al., 2009). Due to the unavailability of a proper documentation of the theory regarding multivariate multilevel binary probit models, it was discussed based on the theory regarding multivariate multilevel probit models for the ordered categorical responses, given by Grilli and Rampichini (2003).

Simple Probit Regression Model Suppose the response of interest which is known as Y can take values 1 and 0 where 1 = higher risk, 0 = lower risk and x can be denoted as set of explanatory variables.

$$\Pr(Y=1|x)=F\left(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k\right), \quad (1)$$

where $F(.)$ is a function such that $F: x \rightarrow [0, 1]$, for all x that belongs to the real line.

The probit model assumes that the function $F(.)$ follows a Normal (cumulative) distribution,

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \varphi(z)\, dz\,, \tag{2}$$

where, $\varphi(z)$ is the Standard Normal Density Function.

$$\varphi(z) = \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi}} \tag{3}$$

Multivariate Multilevel Probit Regression Model Let $Y_{ij}^{(h)}$ be the h^{th} ($h = 1, 2, \ldots, H$) observed binary variable for the i^{th} ($i = 1, 2, \ldots, I$) observation of the j^{th} ($j = 1, 2, \ldots, J$) cluster. Assume that each of the observed responses $Y_{ij}^{(h)}$, which takes values in {1, 2} (for the sake of simplicity, assume it as C) is generated by a latent variable $\tilde{Y}_{ij}^{(h)}$ through the following relationship:

$$\left\{Y_{ij}^{(h)} = C^{(h)}\right\} \text{ if and only if } \left\{\gamma_{C^{(h)}-1}^{(h)} < \tilde{Y}_{ij}^{(h)} < \gamma_{C^{(h)}}^{(h)}\right\} \tag{4}$$

where the threshold satisfies $-\infty = \gamma_0^{(h)} \leq \gamma_1^{(h)} = +\infty$ and γ represents the corresponding value of the response variable when h and c takes values as in equation (4).

Now, consider the following multivariate two-level null model for the latent variables:

$$\tilde{Y}_{ij}^{(h)} = \alpha^{(h)} + u_j^{(h)} + \varepsilon_{ij}^{(h)}, h = 1,..., H\,, \tag{5}$$

where for each h, $\alpha^{(h)}$ is the mean, $u_j^{(h)}$ is the cluster's random effect (level two error) and $\varepsilon_{ij}^{(h)}$ is the individual's disturbance (level one error). The errors are assumed to be distributed as

$$\left[\varepsilon_{ij}^{(1)},...,\varepsilon_{ij}^{(H)}\right]' \sim iidN\left(0,\Sigma_{\epsilon}\right) \text{ and } \left[u_j^{(1)},...,u_j^{(H)}\right]' \sim iidN\left(0,\Omega\right) \quad (6)$$

For example, for $H = 2$ the covariance matrices are,

$$\Sigma_{\epsilon} = \begin{pmatrix} \sigma_{\epsilon 1}^2 & \sigma_{\epsilon 1\epsilon 2} \\ \sigma_{\epsilon 1\epsilon 2} & \sigma_{\epsilon 2}^2 \end{pmatrix}, \Omega = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \quad (7)$$

Moreover, the first and second level errors are assumed to be independent. The previous model specification implies the following conditional covariance structure for any couple of latent variables $\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{ij}^{(k)}\right]$:

$$Cov\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j'}^{(k)} \mid u_j^{(h)},u_{j'}^{(k)}\right] = E\left[\varepsilon_{ij}^{(h)}\varepsilon_{i'j'}^{(k)}\right] = \left\{ \begin{array}{ll} \sigma_{\epsilon_h}^2 & \text{if } k = h, j = j', i = i' \\ \sigma_{\varepsilon_h\varepsilon_k} & \text{if } k \neq h, j = j', i = i' \\ 0 & \text{otherwise} \end{array} \right\}. \quad (8)$$

The unconditional covariance structure is

$$Cov\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j'}^{(k)}\right] = E\left[\varepsilon_{ij}^{(h)}\varepsilon_{i'j'}^{(k)}\right] + E\left[u_j^{(h)}u_{j'}^{(k)}\right], \quad (9)$$

with $Cov\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j'}^{(k)}\right] = 0$ if $j \neq j'$.

The correlation between the same variable for two distinct individuals of the same cluster, called the Intra Cluster Correlation Coefficient (ICC), is stated below.

$$Corr\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j}^{(k)}\right] = \tau_h^2\left(\sigma_{\epsilon_h}^2 + \tau_h^2\right), h = 1,...,H. \quad (10)$$

ICC also represents the proportion of variance explained by the clusters.

Variable selection and model comparison

Consider a subset of covariate and cofactors from the pre-specified set of variables, which best describes the dependent variables. A Forward Selection

procedure along with the Wald Statistic is specifically used for this purpose. MLwiN 2.10 does not use maximum likelihood estimation for estimating the parameters because it is computationally difficult. Therefore as a solution to that, the quasi-likelihood methods were implemented. This shows the inability of considering usual likelihood ratio tests for comparing models.

These methods are implemented by transforming discrete response model to a continuous response model based on a Taylor series expansion. Then, the model becomes linear and then estimation is carried out using Iterative Generalized Least Squares (IGLS) or Reweighted IGLS (RIGLS). These transformations require an approximation known as Marginal Quasi-Likelihood (MQL) and Predictive Quasi-Likelihood (PQL) and can be comprised with Taylor series expansions of either first order terms or second order terms. However, when the sample sizes within level 2 units are small, the first order MQL procedures may lead to biased estimates (Rasbash et al., 2009). Therefore, the second order PQL procedure was adopted.

However due, to some convergence and stability problems it was followed by a Markov Chain Monte Carlo (MCMC) method, which is an alternative to likelihood based estimation procedure.

Parameter Interpretation for a multivariate multilevel model with a probit link function

Interpretation of the coefficients in probit regression is not as straightforward as other regressions. The increase in probability for a unit increment in a given predictor depends on both the values of the other predictors and the initial value of the given predictors. Because the final model presents a lot of interactions and deals with multivariate data in a hierarchical nature, this procedure is more complex and time consuming. Therefore, the probability differences for unit increase of covariates when the other continuous covariates are at their average levels and the categorical covariates are at their base level were considered. Due to the inconvenience of calculating the corresponding probability differences in manual form, a SAS program was used for this purpose.

Residual Analysis and Model Adequacy

It is essential to assess the appropriateness of the fitted model by evaluating the adequacy. Because handling data in a multivariate multilevel framework is a novel approach, diagnostic techniques specifically designed for this scenario are less available. Though the specifications of the models are different according to

the types of variables, the methods of residual analysis and model adequacy are common to all models in a hierarchical nature. Rasbash et al. (2009) presented the theory regarding a basis model having a continuous response in a multilevel data.

Multiple Imputation

The original dataset has small number of observations, removing the records with missing data may cause to create a rather small dataset and it leads to exclude approximately 38% of records. This would cause biased results, because there is a high chance of excluding the low- and middle-income countries from the analysis due to the unavailability of proper information systems. Furthermore, the Missing Data Mechanism (MDM) of this dataset takes the form of Missing At Random (MAR) (Rubin, 1976) and it happens when the missingness depends on a specific variable, but not the value of the variable including missing data (Howell, 2012). In the current dataset, low- and middle-income countries might be less inclined to report their health information due to the unavailability of proper information systems. Therefore, the probability of reported health status is unrelated to the level of health within these low- and middle-income countries and hence the data can be considered MAR. Accordingly, MI was carried to this dataset by using REALCOM (Carpenter et al., 2011) software.

Results

There are three hierarchical levels. Level one consists of the multivariate structure. Level two consists of countries and level three consists of continents. There are 10 variables in the dataset; the countries are clustered within continents. The dataset consist of two response variables termed as CDD and CRC and eight continuous explanatory variables at the country level: Water, Sanitation, Solid_Fuel, B_Glucose, B_Pressure, Obese, Alcohol and Smoking. To implement the univariate analysis these variables were discretized. Although originally there were 195 countries in the dataset, after removing some observations with many missing values and then performing imputation techniques to the variables Smoking, Water and Sanitation, that number was reduced to 186.

Compiled in Table 2 are the *p*-values of the univariate analysis for the associations between imputed explanatory variables with the two outcome variables and the composite variable before and after imputation. The test carried out was the GCMH test. Continent to which the countries belong is used as the second level variable to stratify data accordingly.

Table 2. T_P statistic test results for imputed variables with the responses.

Disease	Explanatory variable	Before			After		
		T_P	DF	*p*-value	T_P	DF	*p*-value
CDD	*Water*	21.888	4	0.00021	23.669	4	9.30E-05
	Sanitation	8.250	4	0.08300	9.954	4	0.04120
	Smoking	8.791	6	0.18600	7.244	6	0.29900
CRC	*Water*	21.302	4	0.00028	21.752	4	0.00022
	Sanitation	23.544	4	9.85E-05	25.307	4	4.36E-05
	Smoking	9.805	6	0.13300	11.118	6	0.08500
CDC+CRC	*Water*	27.201	6	0.00013	27.422	6	0.00012
	Sanitation	19.810	6	0.00299	21.633	6	0.00141
	Smoking	16.869	9	0.05080	11.812	9	0.22410

Note: Consider 20% level of significance

As noted in Table 2, the variables that were considered to be insignificant before imputation remained to be so while those that were significant before imputation remained to be significant apart from the variable Smoking coming under CDD and CDD+CRC which was significant before imputation but had become insignificant after imputation.

Univariate analysis for identifying country level factor impact on the response

Because of the stratified nature of the data, GCMH test was used with a liberal significance level of 20% as explained in Collett (1991). This significance level can be increased because more severe significance levels can lead to the exclusion of potentially useful predictor variables. The requisite calculations were performed using the R-macro developed by De Silva and Sooriyarachchi (2012). Prior to implementing GCMH test, the correlation between CDD and CRC was identified using Pearson's correlation test. For that, two diseases were taken as their continuous form. According to the correlation matrix, there is a significant positive correlation (0.680) that exists between CDD and CRC. Therefore, it is appropriate to perform the Multivariate Multilevel analysis on CDD and CRC.

As noted in Table 3, the two diseases were split into binary outcomes in order to maintain the simplicity of the analysis. Otherwise the resulting composite outcome might have large number of categories and it would be more complex to proceed. The categorization was done by considering the cut-points of worldwide

mortality rates for the two diseases together with the aid of specialists in the field of medicine (World Life Expectancy, n.d.).

Table 3. Categorization of the diseases and description of combined levels

Code (Category)		**Coding for Composite outcomes**
CDD	**CRC**	
1 (<300)	1 (<30)	1
1 (<300)	2 (≥30)	2
2 (≥300)	1 (<30)	3
2 (≥300)	2 (≥30)	4

Compiled in Table 4 are the results of the univariate test, which was carried out to check the significance of country level covariates in the presence of continent as the respective stratification factor for the composite outcome of CDD and CRC.

Table 4. Test Results for composite variable of two diseases vs. Risk Factors

Risk Factors	T_P	**DF**	***p*-value**
Water	27.201	6	0.00013
Sanitation	19.810	6	0.00299
Solid_Fuel	31.403	6	2.10E-05
B_Glucose	14.572	6	0.02390
B_Pressure	21.195	6	0.00170
Obese	19.385	6	0.00356
Alcohol	15.797	9	0.07120
Smoking	16.869	9	0.05080

All the risk factors are significant at a liberal 20% level and the variable Solid Fuel shows the most significance. It implies that there is a higher tendency of getting the disease for the people who are using solid fuel for their day-to-day work.

Fitting a multivariate multilevel probit regression model

Before applying the modeling techniques two diseases were categorized into binary splits as in Table 3. Water and Sanitation were taken as ordered categorical variables while others were taken as their original continuous form. For the multivariate multilevel analysis, there are two types of parameter estimates named

as separate coefficients and common coefficients. Due to that, the model building procedure would be more complex and cumbersome. Therefore, several methods were adopted for the simplification and to obtain an adequate model. For the estimation, the 1st order MQL method was followed by the 2nd order PQL method. It was again followed by the MCMC method to obtain Wald statistic values.

Results indicated that improved drinking water sources and improved sanitation may lead to decrease the incidence of both diseases. This means that the incidence of diseases is increasing when the quality level of water and sanitation are decreasing. Therefore, it would be more meaningful and practicable to get the highest level as the reference for both water and sanitation. Presented in Table 5 are the cofactors and their respective base categories used in the modeling phase.

Table 5. Variables and corresponding base categories

Cofactors	Base category
Water	≥98%
Sanitation	≥80%

At the 1st stage, each factor/covariate was fitted separately and the corresponding Wald statistic value was computed. The *p*-value of the statistic was then compared with the 5% significance level to assess the significance of the coefficient. However, because of the Deviance Information Criteria (DIC) is not available in the MLwiN for the multivariate multilevel scenario, the model building procedure was solely based on the Wald statistic. Forward selection procedure was implemented to identify the main effects. If Wald statistic values for separate coefficients are quite close, the common coefficients should be used as parameter estimates. This argument was used for selecting the other terms as common or separate.

At the 2nd stage each interaction term was fitted separately to the final main effects model. Because there are many interactions pertaining to the variables, fitting all would be more cumbersome and MLwiN would not respond to most of them. Therefore, only the interactions which were significant for the two univariate binomial multilevel logistic regressions for CDD and CRC were considered. However, because there were separate and common coefficients, the interactions were added according to the final main effect model. For an example, consider the B_Glucose*Alcohol interaction. In the final model B_Glucose and Alcohol were fitted as a common coefficient. This means to fit the

B_Glucose*Alcohol interaction also as common coefficients. Figure 1 represents the output of the final interaction model.

$$
\begin{aligned}
&\text{resp}_{1jk} \sim \text{Binomial}(\text{n}_{1jk}, \pi_{1jk})\\
&\text{resp}_{2jk} \sim \text{Binomial}(\text{n}_{2jk}, \pi_{2jk})\\
&\text{resp}^*_{1jk} \sim \text{N}(XB, \Omega)\\
&\text{resp}^*_{2jk} \sim \text{N}(XB, \Omega)\\
&\text{probit}(\pi_{1jk}) = 0.204(0.028)\text{B\_Pressure.CDD}_{ijk} + 0.341(0.450)\text{Sanitation\_1.CDD}_{ijk} +\\
&\qquad 0.407(0.390)\text{Sanitation\_2.CDD}_{ijk} + -0.043(0.023)\text{Obese.CDD}_{ijk} +\\
&\qquad 0.002(0.001)\text{B\_Pressure.Obese.CDD}_{ijk} + -0.087(0.043)\text{Smoking.CDD}_{ijk} +\\
&\qquad 0.003(0.001)\text{B\_Pressure.Smoking.CDD}_{ijk} + u_{0jk}\text{bcons.1}_{ijk} + h_{jk}\\
&\text{probit}(\pi_{2jk}) = 0.232(0.024)\text{B\_Pressure.CRC}_{ijk} + 1.766(0.593)\text{Sanitation\_1.CRC}_{ijk} +\\
&\qquad 1.008(0.387)\text{Sanitation\_2.CRC}_{ijk} + -0.252(0.044)\text{Obese.CRC}_{ijk} +\\
&\qquad 0.008(0.001)\text{B\_Pressure.Obese.CRC}_{ijk} + 0.192(0.046)\text{Smoking.CRC}_{ijk} +\\
&\qquad -0.007(0.001)\text{B\_Pressure.Smoking.CRC}_{ijk} + u_{1jk}\text{bcons.2}_{ijk} + h_{jk}\\
&h_{jk} = \beta_{2k}\text{cons.12} + 1.980(0.462)\text{Water\_1.12}_{jk} + 1.092(0.315)\text{Water\_2.12}_{jk} +\\
&\qquad 0.582(0.039)\text{B\_Glucose.12}_{jk} + -0.387(0.055)\text{Alcohol.12}_{jk} +\\
&\qquad 0.058(0.008)\text{B\_Glucose.Alcohol.12}_{jk} + -0.022(0.001)\text{B\_Glucose.B\_Pressure.12}_{jk} +\\
&\qquad -0.009(0.003)\text{Obese.Alcohol.12}_{jk}\\
&\beta_{2k} = -7.419(0.919) + v_{2k}
\end{aligned}
$$

Figure 1. Final interaction model

According to Figure 1, it can be seen that though there are two levels originally present in the data, the MLwiN is recognized it as three levels. This is because the MLwiN treats the outcomes of two diseases responses as the 1st level (*i*). Therefore $resp_{1jk}$ refers to the number of responses for the disease 1 (CDD) made by the j^{th} country those who are clustered within the continent *k*. Similarly, $resp_{2jk}$ refers to the number of responses for the disease 2 (CRC) made by the j^{th} country those who are clustered within the continent *k*. As a result of that, $resp_{ijk}$ can take either zero or one for all countries in the study. Moreover, n_{1jk} and n_{2jk} always take the value 1, because each country always gives a single response.

Continent level variance component analysis

In order to justify the suitability of applying the multilevel concept, it is advisable to first look at the significance of the continent level variance. This can be checked by the following hypotheses

H_0: Continent level residual variance is zero
H_1: Continent level residual variance is not zero

Because zero is not included in the 95% credible interval (0.254, 5.109), H_0 is rejected and concluded that the continent level variance is significant implying that the multilevel approach for the multivariate context is suitable.

Residual analysis of the final model

After fitting the model the model adequacy was checked. For that purpose, Caterpillar plots and Normal probability plots were used. According to the Caterpillar plot in Figure 2, four residuals do not contain zeros in their 95% confidence bands. These imply significant differences from the overall mean predicted by the fixed part from the model. Moreover, it can be seen that two continents show a negative residual deviation while another two show positive deviations. Therefore it is possible to conclude that these four continents contribute to a high continent effect on the mortality rates of CDD and CRC. These four continents are North America, Europe, Asia and Oceania respectively. Figure 4 illustrates these continent variations more clearly. The continents that have a lower risk are symbolized by green, and higher risk are symbolized by red.

It is suggested in Figure 3 the points are approximately through the 45˚axis indicating that the residuals are approximately normally distributed. However, because of the number of residuals is less, it is hard to conclude the assumption of normality by eye inspection and unable to conduct the Anderson-Darling test with the number of residuals less than seven.

Multivariate Multilevel techniques have recently been developed in the field of statistics and its applications and analysis techniques are very rare. Therefore a suitable goodness of fit test has not yet been developed to evaluate the adequacy of the fitted model. Because there are no other techniques available, the model adequacy was solely dependent on the caterpillar plot and the normal plot.

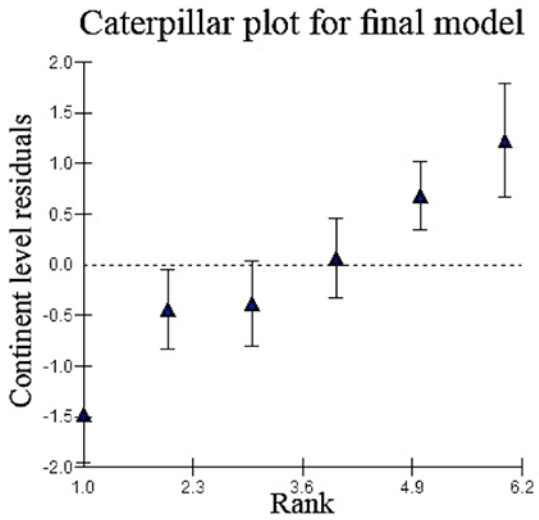


Figure 2. Estimated continent level residuals for the final model

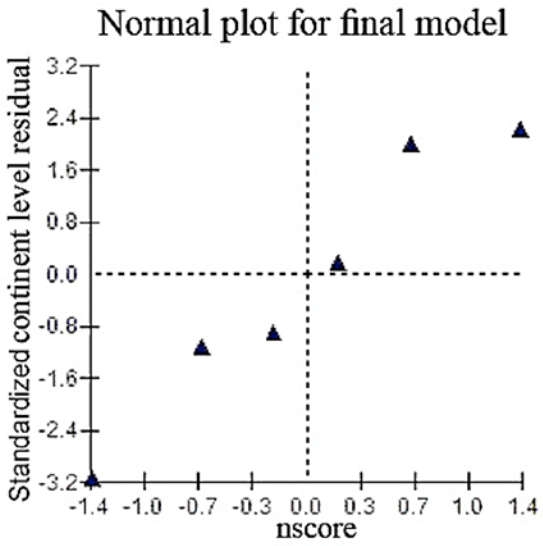


Figure 3. Normal plot for continent level residuals

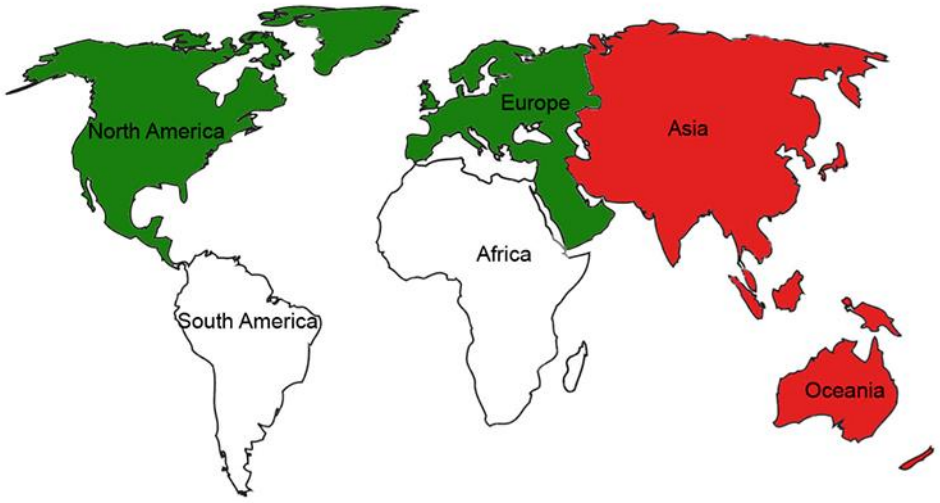


Figure 4. Continent level variations for CDD and CRC

Interpretation and calculation of the parameter estimates

Because the model consists of two equations, due to the multivariate concept this section consists of step-by-step interpretation of each explanatory variable for the two diseases separately. The calculated probability differences are represented in Table 6 and 7.

The results of Table 6 indicate the following important conclusions. The probability of being in the higher group of CDD is 0.6478 higher when Water is at level 1 and 0.3106 higher when Water is at level 2 when compared to level 3 while all the other continuous variables are taken at average and the Sanitation is taken at the base level. However it can be seen that both levels of Sanitation do not have a significant impact for this scenario.

B_Pressure has common interactions with Obese, Smoking and B_Glucose. The probability of CDD being in the higher level compared to the lower level is 0.0149 times more when B_Pressure is increased by one unit and all other variables are at an average and water and sanitation are at base levels.

According to the available medical literature (What Are the Health Risks of Overweight and Obesity?, 2012), it was found that Obesity has shown a higher impact on CDD together with the B_Pressure rather than individually. Therefore, it is more meaningful to identify the combined effect of B_Pressure and Obesity to CDD. The results shows the probability of being in the higher CDD category

compared to the lower CDD category is 0.0081 higher when B_Pressure and Obesity are both increased by one unit when all continuous variables at average and Water and Sanitation are at the base levels.

Table 6. Probability differences for CDD

Term	Probability difference
[a]Water 1	0.6478
[a]Water 2	0.3106
[b]Sanitation 1	Not significant
[b]Sanitation 2	Not significant
[c]B_Pressure = z + 1	0.0149
[c]Obese = y + 1, B_Pressure = z + 1	0.0081
[c]Smoking = s + 1	0.0012
[c]B_Glucose = x + 1	0.0530
[c]Alcohol = w + 1	0.0056

Note: All terms assume continuous variables at average; a) assumes Sanitation = base level; b) assumes Water = base level; c) assumes Sanitation, Water = base level

Table 7. Probability differences for CRC

Term	Probability difference
[a]Water 1	0.5745
[a]Water 2	0.2344
[b]Sanitation 1	0.4911
[b]Sanitation 2	0.2067
[c]B_Pressure = z + 1, Smoking = s + 1	0.0457
[c]Obese = y + 1	-0.0056
[c]B_Glucose = x + 1	0.0341

Note: All terms assume continuous variables at average; a) assumes Sanitation = base level; b) assumes Water = base level; c) assumes Sanitation, Water = base level

Similarly, the probability of CDD being in the higher level compared to the lower level is 0.0012 times higher when Smoking is increased by one unit, 0.053 times more when B_Glucose is increased by one unit and 0.056 times more when Alcohol is increased by one unit while all other variables are at an average and water and sanitation are at base levels

For CRC, the probability of being in the higher group is 0.5745 more when Water is at level 1 and 0.2344 more when it is at level 2 when compared to level 3 while all the other continuous variables are taken at the average and the Sanitation is taken at the base level. Similar to Water, the probability of being in the higher

group of CRC is 0.4911 more when Sanitation is at level 1 and 0.2067 more when Sanitation is at level 2 when compared to level 3. Therefore, it can be seen that when the usage of improved Water and Sanitation sources decreases, the probability of being the higher group of CRC increases.

Smoking has a higher impact to CRC together with B_Pressure rather than individually (Kenny, n.d.). Therefore, when considering the combined effect of those two, the probability of being in the higher CRC category compared to the lower one is 0.0457 times more when both B_Pressure and Smoking are increased by one unit while all continuous variables are at average and Water and Sanitation are at the base levels.

Similarly, the probability of CRC being in the higher level compared to the lower level is 0.0341 times more when B_Glucose is increased by one unit, 0.0693 times more when Alcohol is increased by one unit and 0.0056 times lower when Obesity is increased by one unit while other variables are at an average and water and sanitation are at base levels. Though the latter result seems to be contradictory, it is not so as past medical evidence has suggested that thin people are more prone to get CRC than fat people (Schols et al., 1998).

Discussion

When the usage of unimproved water sources increases, the probability of occurrence of deaths for CDD and CRC also increases. Past evidence also indicated this relationship. Fodor et al. (1973) showed the proportion of mortality rates for CDD was higher in the soft water areas than hard water areas. It was further shown there was a macro geography variation for CDD. Those findings tally with the findings in this study because here also CDD shows a continent level variation. They also showed CRC has an impact from the variable Water. But it is a less known thing. However, officials at the US Environmental Protected Agency suggested heavy rainfall events cause storm water overflow that may contaminate water bodies used for drinking with other bacteria. It may cause to get illnesses, including ear, nose, and throat infections (Climate impacts on Human Health, n.d.).

Although CDD has no impact from Sanitation, the probability of being in the higher group of CRC increases due to the usage of unimproved sanitation sources. When analyzing risk factors for diseases, the focus is less given for the environmental factors such as water, sanitation etc. However, it was shown the usage of unimproved Water and Sanitation sources have more impact to the diseases CDD and CRC. According to Briggs's (2003), unsafe water, poor

sanitation and poor hygiene seem to be one of the major sources of exposure for these types of diseases.

National Heart, Lung and Blood Institute (What Are the Health Risks of Overweight and Obesity?, 2012) claimed most people who have type 2 diabetes are overweight and also it leads to heart failures. Furthermore, they have shown that the chances of having high blood pressure are greater if people are overweight. This joint impact of B_Pressure and Obesity on CDD by showing the probability of occurrence of death in CDD increases when both B_Pressure and Obesity are increased by one unit.

When considering CRC, medical evidence (Kenny, n.d.) suggests that chronic obstructive pulmonary disease (COPD) usually cause by smoking and continuous smoking for a long time causes to increase breathing difficulties and also causes to increase blood pressure. As a result of that it can put a heavy strain on the heart muscle and creates heart failures. After that Respiratory failures occur as the final stage of COPD (Kenny, n.d.). This factor shows that there is an interesting flow by beginning from smoking through the increment of blood pressure to the respiratory failures. This further demonstrates an interesting relationship by showing increase in the probability of being in the higher level of CRC compares to the lower level when B_Pressure and Smoking are both increased by one unit.

Some medical evidence (Schols et al., 1998) suggested it is difficult to identify a suitable relationship between Obesity and CRC. A decrement of the probability being in the higher level of CRC for a unit increment of Obesity was shown. However, a large epidemiologic study showed overweight and obesity in patients with COPD was associated with a decreased risk of death compared with normal weight (Schols et al., 1998). Therefore, it might be concluded that thin people are more prone to get CRC than obese people. Furthermore, North America and Europe show a less risk of having CDD and CRC while Asia and Oceania show a higher risk with CDD varies less with continent while CRC varies more.

Limitations of the study

In the advanced analysis phase, logistic and probit regression models were fitted with the continuous explanatory variables. The models contained common interactions as well as cross interactions. Therefore, it was more complex to obtain corresponding confidence intervals for the odds ratios and for the predicted

probabilities and hence the significance/non-significance of the estimates could not be evaluated.

The interpretations of coefficients in multivariate multilevel binary probit regression models are not as simple as in other models (i.e., linear regression, logit regression, etc.). Increment in probability for a unit increment in a given predictor depends both on the values of the other predictors and the initial value of the given predictors. Therefore, the results can be changed due to the different values of the predictors.

Of note, in the advanced model building phase, MLwiN crashed many times, therefore some of the terms had to be excluded from the initial model. This might have happened due to small number of data points and non-convergence of models.

Conclusion

In Multivariate multilevel model building process there is no satisfactory goodness of fit test yet developed. Therefore it is essential to develop a goodness of fit test in order to access the model adequacy of the multivariate multilevel models. The mortality rates of Asia and Oceania should be reduced, by improving health policies to meet standards like those in North America and Europe. Furthermore, higher consideration should be given to environmental risk factors such as water quality and sanitation to improve personal health.

References

Briggs, D. (2003). Environmental pollution and the global burden of disease. *British Medical Bulletin, 68*(1), 1-24. doi: 10.1093/bmb/ldg019

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software, 45*(5). doi: 10.18637/jss.v045.i05

Climate impacts on Human Health. (n.d.). Retrieved from http://www.epa.gov/climatechange/impacts-adaptation/health.html, Accessed 4 December 2014.

Collett, D. (1991). *Modelling binary data* (1st ed.). London: Chapman & Hall. doi: 10.1007/978-1-4899-4475-7

De Silva, D. & Sooriyarachchi, M. (2012). Generalized Cochran Mantel Haenszel test for multilevel correlated categorical data: an algorithm and R

function. *Journal of the National Science Foundation of Sri Lanka, 40*(2), 137-148. doi: 10.4038/jnsfsr.v40i2.4441

Fodor, J. G., Pfeiffer, C. J. & Papezik, V. S. (1973). Relationship of drinking water quality (hardness-softness) to cardiovascular mortality in Newfoundland. *Canadian Medical Association Journal 108*(11), 1369–1373.

Grilli, L. & Rampichini, C. (2003). Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics 28*(1), 31-44. doi: 10.3102/10769986028001031

Howell, D. C. (2012). Treatment of Missing Data-Part 1. Retrieved from http://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html, Accessed 5 November 2014.

Jayawardana, N. I. & Sooriyarachchi, M. R. (2014). A multilevel Bayesian analysis of university entrance eligibility for selected districts in Sri Lanka: methods and application to educational data. *Journal of the National Science Foundation of Sri Lanka, 42*(1), 23-36. doi: http://dx.doi.org/10.4038/jnsfsr.v42i1.6676.

Kenny, T. (n.d.). Chronic obstructive pulmonary disease. Retrieved from http://www.patient.co.uk/health/chronic-obstructive-pulmonary-disease-leaflet, Accessed 4 December 2014.

Rasbash, J., Steele, F., Browne, W. J. & Goldstein, H. (2009). *A user's guide to MLwiN, v2.10*. Bristol, UK: University of Bristol.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581-592. doi: 10.1093/biomet/63.3.581

Schols, A. M., Slangen, J., Volovic,s L. & Wouters, E. F. (1998). Weight loss is a reversible factor in the prognosis of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine, 157*(6), 1791-1797. doi: 10.1164/ajrccm.157.6.9705017

Snijders, T. A. B. & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). London, UK: Sage Publications Ltd.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ, 338*, b2393. doi: 10.1136/bmj.b2393

What Are the Health Risks of Overweight and Obesity? (13 July, 2012). Retrieved from http://www.nhlbi.nih.gov/health/health-topics/topics/obe/risks, Accessed 2 December 2014.

World Health Organization. (2005). *Preventing chronic diseases: a vital investment.* Retrieved from http://www.who.int/chp/chronic_disease_report/full_report.pdf, Accessed 1 December 2014.

World Health Organization. (2013). *World health statistics 2013*. Retrieved from http://www.who.int/gho/publications/world_health_statistics/2013/en/, Accessed 1 December 2014.

World Life Expectancy. (n.d.) Retrieved from http://www.worldlifeexpectancy.com, Accessed 2 December 2014.

Zhang, J. & Boos, D. D. (1997). Mantel-Haenszel test statistics for correlated binary data. *Biometrics, 53*(4), 1185-1198. doi: 10.2307/2533489

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 518-543.
doi: 10.22237/jmasm/1493598600

ISSN 1538 − 9472

A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys

Si Yang
University of Rhode Island
South Kingstown, RI

Lisa L. Harlow
University of Rhode Island
South Kingstown, RI

Gavino Puggioni
University of Rhode Island
South Kingstown, RI

Colleen A. Redding
University of Rhode Island
South Kingstown, RI

The performance of several models under different conditions of zero-inflation and dispersion are evaluated. Results from simulated and real data showed that the zero-altered or zero-inflated negative binomial model were preferred over others (e.g., ordinary least-squares regression with log-transformed outcome, Poisson model) when data have excessive zeros and over-dispersion.

Keywords: zero-inflated analysis, count data

Introduction

In psychological, social, and public health related research, it is common that the outcomes of interest are relatively infrequent behaviors and phenomena. Data with abundant zeros are especially frequent in research studies when counting the occurrence of certain behavioral events, such as number of school absences, number of cigarettes smoked, number of hospitalizations, or number of unhealthy days. These types of data are called count data and their values are usually non-negative with a lower bound of zero and typically exhibit excessive zeros and over-dispersion (i.e., greater variability than expected).

Except for transforming the outcome to make it normal and using the general linear model, other alternative approaches can be taken in the context of a broader framework: generalized linear model (GLM). For example, the Poisson distribution becomes increasingly positively skewed as the mean of the response

Si Yang is a Statistical Consultant and former Instructor at the University of Rhode Island. Email at yangsi06@gmail.com.

variable decreases, which reflects a common property of count data (Karazsia and Van Dulmen, 2008). Thus, a typical way of analyzing count data includes specification of a Poisson distribution with a log link (the log of the expectation of a response variable is predicted by the linear combination of covariates, i.e., predictors) in a model known as Poisson regression.

Several other more rigorous approaches to analyzing count data include the zero-inflated Poisson (ZIP) model and the zero-altered Poisson model (ZAP, also called a hurdle model) that have been proposed recently to cope with an overabundance of zeros (Greene, 1994; King, 1989; Lambert, 1992; Mullahy, 1986). These two types of models both include a binomial process (modeling zeros versus non-zeros) and a count process. The difference between the two models is how they deal with different types of zeros: although the count process of ZAP is a zero-truncated Poisson (i.e. the distribution of the response variable cannot have a value of zero), the count process of ZIP can produce zeros (Zuur, et al., 2009). One of the assumptions of using Poisson regression is that the mean and variance of a response variable are equal. In reality, it is often the case that the variance is much larger than the mean. Variations of negative binomial (NB) models can be used when over-dispersion exists even in the non-zero part of the distribution. Although a Poisson distribution contains only a mean parameter (μ), a negative binomial distribution has an additional dispersion parameter (k) to capture the amount of over-dispersion. Thus, the zero-inflated negative binomial (ZINB) model and zero-altered negative binomial (ZANB) model were introduced to deal with both zero-inflation and over-dispersion.

Traditionally, dichotomizing or transforming the dependent variables have been used as solutions to handle the non-normality of the data. Approaches such as a Poisson model, NB model, ZIP/ZAP models, or ZINB/ZANB models have recently been demonstrated and compared to analyze zero-inflated count data through several tutorial style papers (e.g., Atkins, 2012; Karazsia and Van Dulmen, 2008; Loeys, et al., 2012; Vives, et al., 2006). Each of these papers largely focus on a single empirical study and models were only being compared in one condition. The current study focused on comparing a set of models under different conditions of zero-inflation and skewness and aimed to offer clear guidelines as to which model to use under a certain condition.

GLM and Poisson regression

The GLM is a flexible modeling framework that allows the response variables to have a distribution form other than normal. It also allows the linear model of

several covariates to be related to a response variable via arbitrary choices of link functions. Zuur et al. (2009) summarized that building a GLM consists of three steps: a) choosing a distribution for the response variable (Y); b) specifying covariates (X); and c) choosing a link function between the mean of the response variable ($E(Y)$) and a linear combination of the covariates (βX). Classical models such as analysis of variance (ANOVA) and ordinary least squares regression also belong to the GLM when Y is normally distributed. Y can also be specified as other distributional forms in the exponential family such as a binomial distribution, Poisson distribution, negative-binomial distribution, and gamma distribution. The link function brings together the expectation of the response variable and the linear combination of the covariates. For ordinary least-squares regression, the function to estimate the expected value of Y is $\beta X = E(Y)$; it is termed as an identity link. Specifying a logit link as $\beta X = \log(E(Y) / (1-E(Y)))$ is usually used for logistic regression to predict the expectation of a binary response variable. The probability mass function (p.m.f) of a Poisson distribution is as follows:

$$\Pr\left(Y_i = y_i\right) = \frac{e^{-m} m^{y_i}}{y_i!},\ y_i = 0,1,2,...$$

where μ is the count mean. Let $X = (X_1, \ldots, X_p)$ be a vector of covariates and $\beta = (\beta_1, \ldots, \beta_p)$ be a vector of regression parameters. The logarithm of μ is assumed to be a linear combination of p covariates of the form

$$E\left(Y \mid X\right) = m = \exp\left(Xb\right)$$

The conditional mean and conditional variance are equal for the Poisson regression model, that is $E(Y|X) = \mathrm{Var}(Y|X) = \mu$. The greater the mean the greater is the variability of the data. A large proportion of zeros in the count data leads to a smaller mean value than that of the variance.

Negative binomial regression model

The assumption that the variance of counts is equal to the mean also implies that the variability of the outcomes sharing the same covariates values (a population has the same values for $X_1, X_2, \ldots, X_p$) is equal to the mean. If it fails to be true, the estimates of the regression coefficients can still be consistent using Poisson regression, but the standard errors can be biased. They usually tend to be too

small and thus increase the rate of Type I error (false positive results) (Hilbe, 2014). When analyzing data to explore relationships between variables or make predictions, we would not expect we have measured every variable that contributes to the rates of the outcome events. There will always be residual variation in the response variables. For instance, Roebuck et al. (2004) studied how adolescent marijuana use might relate to school attendance (estimated by number of days truant) by analyzing data from the National Household Survey on Drug Abuse. It is unlikely that adolescent marijuana users will have the same rate of being truant; specifically, there is more variation in school attendance among marijuana users. To account for greater variation, the negative binomial model has been proposed as a generalization of the Poisson model. The negative binomial distribution has the following form:

$$\Pr\left(Y_i = y_i\right) = \frac{\Gamma\left(k + y_i\right)}{\Gamma\left(k\right) y_i} \left(\frac{k}{k+m}\right)^k \left(\frac{m}{k+m}\right)^{y_i}$$

where μ is the mean and k is the dispersion parameter. The variance of the above distribution is $\mu + \mu^2/k$, and hence decreasing values of k correspond to increasing levels of dispersion. As k increases towards positive infinity, a Poisson distribution is obtained. The negative binomial regression model is able to capture the over-dispersion in count data that the simple Poisson model cannot. However, the problem of excessive zeros is still not solved, as researchers may be interested in finding the special meaning underlying the zero-inflation.

Zero-inflated regression models

Lambert (1992) proposed an approach to model zero-inflation in count data in what is referred to as a ZIP model. In this model, two kinds of zeros are thought to exist in the data: structural zeros (or true zeros) from a non-susceptible group (i.e., those that do not have the attribute or experience of interest, such as healthy people without a disease) and random zeros (or false zeros) for those from a susceptible group (e.g., those that have a disease in a health-based study who may falsely indicate a score of zero). The probability of being in a susceptible group can be estimated by information from covariates with a logistic link. If an individual is from the susceptible group, his or her count is a random variable from a Poisson distribution with mean μ. The marginal distribution of the ZIP model is as follows:

$$\Pr\left(Y_i = y_i\right) = \begin{cases} \left(1 - p\right) + pe^{-m}, & \text{for } y_i = 0 \\ p\dfrac{e^{-m}m^{y_i}}{y_i!}, & \text{for } y_i = 1,2,... \end{cases}$$

The Poisson hurdle model (i.e., ZAP) as an alternative was introduced by Mullahy (1986) and modified by King (1989). It models all zeros as one part and a zero-truncated part for all non-zero observations. The main difference with ZIP is that hurdle models don't distinguish true and false zeros and all zero observations are thought to come from a non-susceptible group:

$$\Pr\left(Y_i = y_i\right) = \begin{cases} 1 - p, & \text{for } y_i = 0 \\ p\dfrac{e^{-m}m^{y_i}}{y_i!\left(1 - e^{-m}\right)}, & \text{for } y_i = 1,2,... \end{cases}$$

Because a Poisson distribution assumes that the variance of the outcome variable equals its mean, when over-dispersion also comes from the non-zero part (i.e., the variance is much bigger than the mean even for the non-zero part), both ZIP and ZAP models can be extended to ZINB or ZANB models to deal with zero-inflation and over-dispersion at the same time. These types of models have become popular recently and have been used to analyze number of cigarettes smoked per day (Schunck & Rogge, 2012), dental health status (Wong & Lam, 2012), depressive symptoms (Beydoun, et al., 2012), and alcohol consumption (Atkins, 2012), etc. The major advantage of using models specially dealing with zero-inflation is that they not only reduce biases resulting from the extreme non-normality but also have the ability to model the effect on subjects' susceptibility and magnitude at the same time.

Proposed Study

For count data, depending on an outcome's mean-variance relationship and proportion of zeros, the choices for modeling its distribution range from standard Poisson and negative binomial to ZIP, and ZINB (or ZAP and ZANB). However, some researchers argue that they have seen cases where ZIP models were inadequate and ZINB also couldn't be reasonably fitted to the data (Famoye & Singh, 2006). Warton (2005) also criticized such zero-inflated models as being too routinely applied, leading to overuse. He analyzed 20 multivariate abundance

datasets extracted from the ecology literature using three different approaches: least squares regression on transformed data, log-linear models (Poisson and negative binomial regression), and zero-inflated models (ZIP and ZINB), and then compared each model's goodness-of-fit. The result showed that a Gaussian (i.e., normally distributed) model (e.g., least squares regression) based on a transformed outcome fit the data surprisingly better than fitting zero-inflated count distributions. This study also suggested that negative binomial regression had the best fit, and that special techniques for dealing with excessive zeros may be unnecessary.

Based on these open questions in the field, there appears to be a conflict since there is increasing popularity of zero-inflated models, although some empirical evidence has tended to show no better fit for these models compared with the traditional least squares method conducted on transformed data. Moreover, there is much disagreement about which zero-inflated model to choose from among ZIP, ZINB, ZAP, and ZANB. In the zero-inflation data analysis literature, proposing an extensional zero-inflated model or comparing different models are often motivated and illustrated by a single empirical study. These can look more like case studies in which each dataset or applied situation has its particular uniqueness. It is possible that the discrepancy in the results from these studies depends on having a different proportion of zeros and different skewness in the non-zero part. It is becoming apparent that having data with excessive zeros is the norm in many situations, with or without known reasons. However, it is not clear what the proportion of zeros is, after which the data should be considered as zero-inflated, and what the underlying mechanism of abundant zeros is. Further, when researchers have collected data with abundant zeros, should zero-inflated models be used, and if so, which one should be used? These are questions that have unclear or controversial answers in the zero-inflation literature, and which are driving the proposed research. This study used systematic methods to try to answer these questions.

Another consideration is that, a full range of these methods hasn't been compared and tested under different conditions. The purpose of this study was to examine the performance of different techniques dealing with zero-inflation. Both simulated data and empirical data with and without known reasons for zero-inflation were analyzed. Specifically, this study addressed the following research questions:

1. Under conditions of different degrees of zero-inflation (i.e. proportion of zeros in the response variable) but the same level of dispersion,

which of the following models is superior: a) least squares regression with a transformed outcome; b) Poisson regression; c) negative binomial regression; d) ZIP; e) ZINB; f) ZAP; or g) ZANB?

2. Under conditions of different degrees of dispersion but the same zero-inflation level, which of the following models is superior: a) least squares regression with a transformed outcome; b) Poisson regression; c) negative binomial regression; d) ZIP; e) ZINB; f) ZAP; or g) ZANB?

3. Finally, for the empirical data from a national health survey with a zero-inflated and over-dispersed response variable, which of the following models is superior: a) least squares regression with a transformed outcome; b) Poisson regression; c) negative binomial regression; d) ZIP; e) ZINB; f) ZAP; or g) ZANB?

Methods

Simulation

Simulation Study Design Data were generated with a mix of zeros and a negative binomial distribution. A brief literature review on the frequency of various health survey outcomes showed that the percentage of zeros tends to range from 20% to 90% (Beydoun, et al., 2012; Lin & Tsai, 2012; Mahalik, et al., 2013); thus, four conditions with varying probability of zeros (w in Table 1) for the response variable were tested in the current study to reflect this range. A condition of no zero-inflation ($w = 0.00$) was also tested as a baseline comparison. In order to examine the effect of over-dispersion in the non-zero part, a dispersion parameter k with the following values: 1, 5, 10, and 50 were pre-specified. These values represent a reasonable range of dispersion to help assess the merit of various models with varying distributions. The bigger the k the less dispersed the variable is and it approaches a Poisson distribution when $k > 10\mu$ (Bolker, 2008). The response variable was generated with a negative binomial distribution with a different proportion of zeros added. The simulation study was a 5 (i.e., Factor A: degree of zero-inflation) x 4 (i.e., Factor B: degree of dispersion) factorial design that was examined for the 7 models listed for Factor C, as shown in Table 1.

Table 1. Simulation design factors

Factor A	***Factor B***	***Factor C***
w	***k***	**Models** (Tested on each of the 5×4 conditions in A & B)
0.00 0.20 0.40 0.60 0.80	1 5 10 50	Least squares regression with transformed outcome (LST) Poisson regression model (Poisson) Negative binomial regression model (NB) Zero-inflated Poisson model (ZIP) Zero-inflated negative binomial model (ZINB) Zero-altered Poisson model (ZAP) Zero-altered negative binomial model (ZANB)

Note. Factor A indicates the proportion of zeros in the simulated data, ranging from w = 0 (i.e., none) to .80 (i.e., high). Factor B indicates the degree of dispersion in the data, ranging from k = 1 (i.e., high) to 50 (i.e., low).

Generating Simulated Datasets To provide a reasonable prediction model to explore in this study, a count response variable Y and two different kinds of covariates, X_1 and X_2, were simulated. X_1 was assumed to be a binary variable whose values were 0 or 1 with $\Pr(X_1 = 0) = \Pr(X_1 = 1) = 0.5$. X_2 was set to follow a standard normal distribution, $N(0,1)$. Regression coefficients β_1 and β_2 for the two covariates were set to be 0.3 and 0.5 for the population model to allow for a medium and large value, respectively. It is recognized that the two values cannot be seen as standardized effect sizes as the scores for Y and X_1 are not standardized. However, regression coefficients of 0.3 and 0.5 can be seen as reasonable choices that allow for a comparison between different levels of prediction for the two covariates. To ensure accurate results, 2000 replications (i.e., simulation size, $S = 2000$), each with sample size $n = 500$, were generated. The simulated mean for the count process (μ) was 1.33 (SD = 0.03) across all simulations. The decisions on the number of simulations and sample size were made by referring to previous simulation studies on zero-inflated data (e.g., Lambert, 1992; Min & Agresti, 2005; Williamson, et al., 2007).

Model Selection Criteria The model with minimum AIC (Akaike information criterion) was considered as the best model to fit the data (Bozdogan, 2000). AIC is given by:

$$\text{AIC} = -2\log L(\theta) + 2c,$$

where $L(\theta)$ is the maximized likelihood function for the estimated model and $-L(\theta)$ offers summary information on how much discrepancy exists between the model and the data, where c is the number of free parameters in the model.

AIC assesses both the goodness of fit of the model and the complexity of the model. It rewards the model fit by the maximized log likelihood term $2\log L(\theta)$, and also prefers a relatively parsimonious model by having c as a measure of complexity. There are two challenges for calculating a comparable AIC for the LST model. First, AIC can only be used to compare models with the exact same response variable. Second, a response variable in the LST model is assumed to be continuous, whereas in other models it is a count. It is not correct to compare the log-likelihood of discrete distribution models and continuous distribution models, as the former is the sum of the log probabilities and the latter is the sum of the log densities. Warton (2005) used a discretization method to address the issue and we applied the same approximation approach in this paper. For the LST model, the Gaussian distribution for AIC calculation was discretized as below.

$$L(\theta) = L(\hat{\mu}, \hat{\sigma}; y) = \sum_{i=1}^{N} \log\left\{ \Phi\left[\frac{\log(y_i + 1.5) - \hat{\mu}_i}{\hat{\sigma}} \right] - \Phi\left[\frac{\log(y_i + 0.5) - \hat{\mu}_i}{\hat{\sigma}} \right] \right\},$$

where $\hat{\mu}_i$ and $\hat{\sigma}$ are the estimated mean and standard deviation of the response varaible y, and $\Phi(c)$ is the lower tail probability at c from the standard normal distribution.

Empirical Data Analysis

Analyses were conducted on an existing data set to further assess different procedures. The Behavioral Risk Factor Surveillance System (BRFSS) collected information on health risk behaviors, health conditions, health care access, and use of preventive services (CDC, 2012). In this portion of the study based on actual data, the relationship between physical activity and health related quality of life was examined after controlling for age and gender, continuous and binary covariates, respectively.

Participants The data were obtained from the 2011 Rhode Island BRFSS, a random-digit telephone health survey of adults 18 years of age or older. Of 6533 participants involved in the survey, 38.3% were males and 61.7% were females ranging in age from 18 to 98 (M = 55.51, SD = 16.90).

Measures

Health Related Quality of Life (HRQoL): The overall number of mentally or physically unhealthy days (UNHLTH) in the last 30 days was used as an indicator of having poor HRQoL. The summary index of unhealthy days was calculated by combining the following two questions (CDC, 2012), with a logical maximum of 30 unhealthy days:

1. "Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?"
2. "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?"

Physical Activity (PA): A set of questions in the BRFSS captured data on three key domains of physical activity: leisure-time, domestic, and transportation. A summary score for physical activity was calculated and then was categorized into four levels according to CDC's 2008 Physical Activity Guidelines for Americans, a) highly active, b) active, c) insufficiently active, and d) inactive, with higher scores indicating higher levels of physical activity.

Analysis Participants reporting 30 physically or mentally unhealthy days during the past month were not included in the analysis. These individuals were considered as patients with long-term sickness who did not meet the inclusion criteria for this study. PA, age, gender, and their interactions with PA were entered as predictors of having poor HRQoL. Seven regression models described above were used to fit the data. In addition to using AIC values to evaluate the models, Vuong's tests were also used for model comparisons. Vuong's test is likelihood-ratio based for comparing nested, non-nested, or overlapping models in a hypothesis testing framework (Vuong, 1989). The null hypothesis was that both models were equally close to the true model. To control for Type I error rate for the several model comparisons that were made, $p < .01$ was used as a criterion for a statistically significant result.

Statistical Program R (R Core Team, 2013) was used for both data simulation and data analyses. Function `rnbinom()` was used to generate random negative binomial variables. Functions `hurdle()` and `zeroinfl()` from package

`pscl` (Jackman, 2008) were used to fit data with zero-altered and zero-inflated models; and `glm()` from package `stats` was used to fit LST, Poisson, and NB models.

Results

Results from simulation study

Average AIC values and selection rates (i.e., percentages of runs having the lowest AIC, which indicated a more preferred model) across all simulations for the five levels of zero-inflation combined with four levels of over-dispersion on the seven models are presented in Table 2.

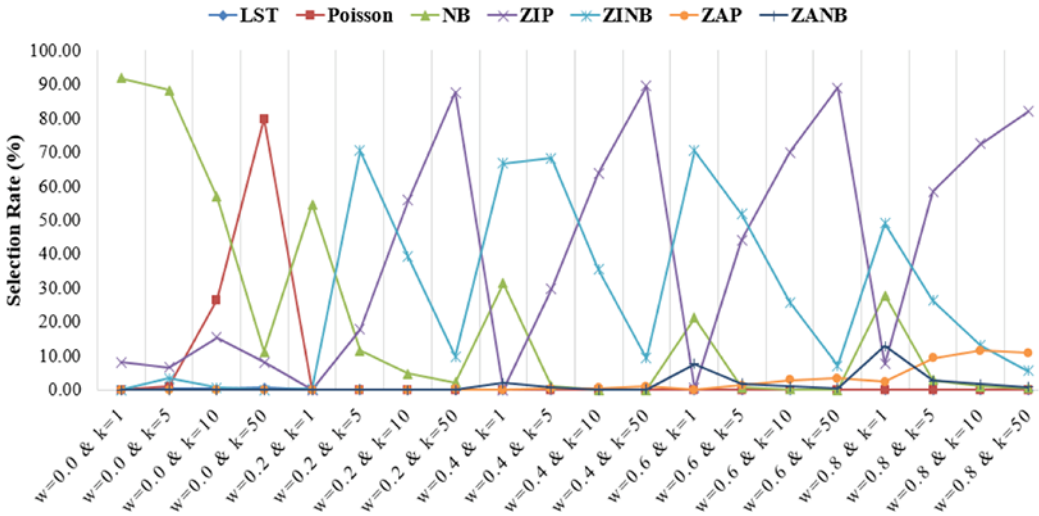
Figure 1 gives a visual presentation of how selection rates changed across different conditions for different models. Under the no zero-inflation condition (w = 0.0), a Poisson model was more preferred when k = 50 (i.e., low dispersion) and a NB model was more preferred when k = 1, 5, or 10 (i.e., high to moderate dispersion). When data did exhibit zero-inflation, even with just 20% of zeros, a ZIP model was more preferred with low dispersion (k = 10 or 50); a ZINB model was more preferred with high dispersion (k = 1 or 5); the Poisson model and the LST model yielded much larger average AIC values with a 0% selection rate; and the NB model had higher selection rates as k and w got smaller (i.e., high dispersion and low proportion of zeros). The ZIP, ZINB, ZAP, and ZANB had similar AIC values across all of the conditions, however, ZIP and ZINB had much higher percentages of being more preferred models compared with ZAP and ZANB.

Boxplots for the AIC values across different conditions were constructed for the seven models. Figures 2.1 and 2.5 show the most (k = 1) and least (k = 50) over-dispersed levels of the five conditions of proportion of zeros (i.e., w = 0.0, 0.2, 0.4, 0.6, and 0.8). For each figure, the left side pertains to k = 1 and the right side to k = 50. Further, a reference line was added to all figures by using the minimum mean AIC values. For definitions of the seven models, refer to the note in Figure 1.

Table 2. Mean AIC values, and percentage with the lowest AIC across all simulations (in parenthesis), for 12 conditions on 7 models

Conditions		LST	Poisson	NB	ZIP	ZINB	ZAP	ZANB
w = 0.0	*k = 1*	1579.19 (0.00)	1724.70 (0.00)	1521.05 **(91.80)**	1603.99 (8.15)	1522.51 (0.00)	1630.84 (0.00)	1553.50 (0.05)
	k = 5	1476.20 (0.50)	1471.66 (1.00)	1456.48 **(88.35)**	1465.11 (6.70)	1457.99 (3.45)	1520.47 (0.00)	1513.84 0.00
	k = 10	1450.21 (0.45)	1435.32 (26.55)	1432.23 **(56.80)**	1434.54 (15.45)	1433.73 (0.75)	1496.19 (0.00)	1495.29 (0.00)
	k = 50	1425.32 (0.75)	1406.15 **(79.85)**	1407.34 (11.30)	1407.50 (8.10)	1409.03 (0.00)	1474.36 (0.00)	1475.72 (0.00)
w = 0.2	*k = 1*	1457.22 (0.00)	1615.49 (0.00)	1354.40 **(54.60)**	1416.87 (0.00)	1353.76 (0.35)	1433.80 (0.00)	1373.51 (0.00)
	k = 5	1407.79 (0.00)	1416.70 (0.00)	1358.24 (11.55)	1358.28 (17.80)	1352.76 **(70.50)**	1389.93 (0.00)	1384.92 (0.15)
	k = 10	1392.36 (0.00)	1384.38 (0.00)	1348.08 (4.80)	1340.95 **(55.95)**	1340.27 (39.15)	1375.28 (0.00)	1374.78 (0.10)
	k = 50	1382.03 (0.00)	1363.17 (0.00)	1340.78 (2.25)	1329.22 **(87.65)**	1330.53 (9.95)	1365.27 (0.15)	1366.62 0.00
w = 0.4	*k = 1*	1292.70 (0.00)	1435.58 (0.00)	1135.39 (31.35)	1178.50 (0.00)	1132.75 **(66.65)**	1189.51 (0.00)	1145.75 (2.00)
	k = 5	1271.47 (0.00)	1290.11 (0.00)	1178.62 (1.15)	1170.28 (29.65)	1166.76 **(68.25)**	1189.09 (0.30)	1185.91 (0.65)
	k = 10	1266.32 (0.00)	1269.65 (0.00)	1182.15 (0.10)	1166.74 **(63.80)**	1166.68 (35.50)	1186.98 (0.55)	1187.06 (0.05)
	k = 50	1257.74 (0.00)	1249.31 (0.00)	1179.71 (0.05)	1159.01 **(89.40)**	1160.42 (9.40)	1180.13 (1.00)	1181.58 (0.15)
w = 0.6	*k = 1*	1078.86 (0.00)	1171.71 (0.00)	861.25 (21.30)	886.33 (0.50)	857.62 **(70.50)**	892.43 (0.10)	864.80 (7.60)
	k = 5	1071.22 (0.00)	1075.19 (0.00)	920.11 (0.70)	908.89 (44.20)	907.18 **(51.75)**	919.48 (1.45)	918.02 (1.90)
	k = 10	1067.62 (0.00)	1060.84 (0.00)	925.78 (0.15)	909.23 **(69.90)**	909.59 (25.90)	920.30 (2.95)	920.77 (1.10)
	k = 50	1063.87 (0.00)	1047.59 (0.00)	931.34 (0.00)	910.16 **(89.00)**	911.68 (7.10)	921.81 (3.35)	923.36 (0.55)
w = 0.8	*k = 1*	782.26 (0.00)	765.93 (0.00)	516.17 (27.90)	525.66 (7.65)	513.55 **(49.15)**	528.35 (2.40)	516.84 (12.90)
	k = 5	775.82 (0.00)	720.75 (0.00)	563.92 (2.95)	555.29 **(58.45)**	555.32 (26.40)	559.70 (9.40)	559.88 (2.80)
	k = 10	773.28 (0.00)	712.79 (0.00)	571.38 (1.00)	559.97 **(72.45)**	561.04 (13.15)	564.58 (11.60)	565.73 (1.80)
	k = 50	772.36 (0.00)	708.09 (0.00)	576.99 (0.55)	563.21 **(82.05)**	564.79 (5.65)	568.29 (10.85)	569.91 (0.90)

Note: Numbers in parentheses are percentages (%) of simulations out of 2,000 simulations in which model had the lowest AIC value (most preferred); *w* is the proportion of zeros and *k* is the dispersion parameter used to simulate the data. LST = least squares regression with transformed outcome, Poisson = Poisson regression model, NB = negative binomial regression model, ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, ZAP = zero-altered Poisson model, ZANB = zero-altered negative binomial model.



Note: *w* is the proportion of zeros and *k* is the dispersion parameter used to simulate the data. LST = least squares regression with transformed outcome, Poisson = Poisson regression model, NB = negative binomial regression model, ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, ZAP = zero-altered Poisson model, ZANB = zero-altered negative binomial model.

Figure 1. Percentages of having the lowest AIC across 2000 simulations

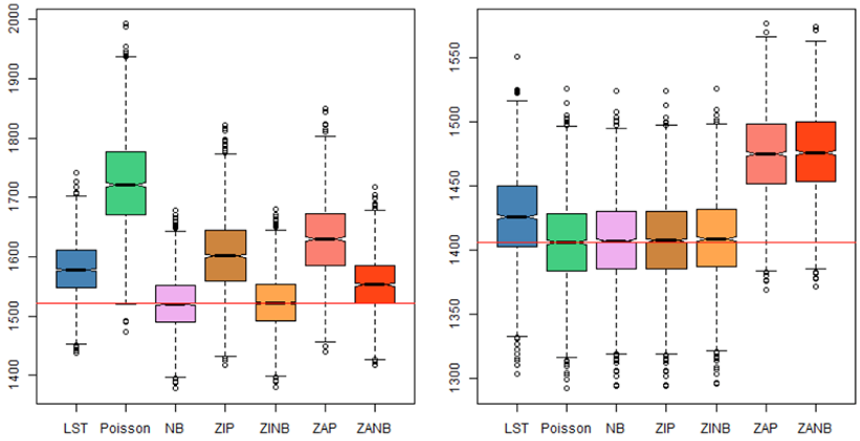


Figure 2.1. Boxplot of AIC from seven models (*w* = 0.0)

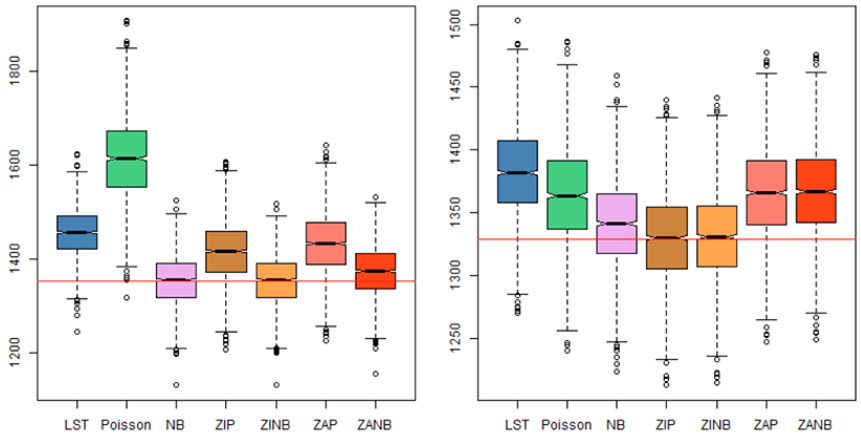


Figure 2.2. Boxplot of AIC from seven models (w = 0.2)

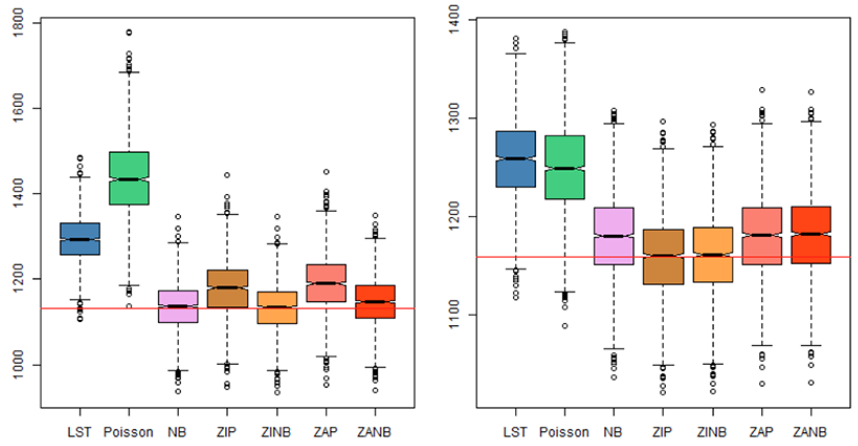


Figure 2.3. Boxplot of AIC from seven models (w = 0.4)

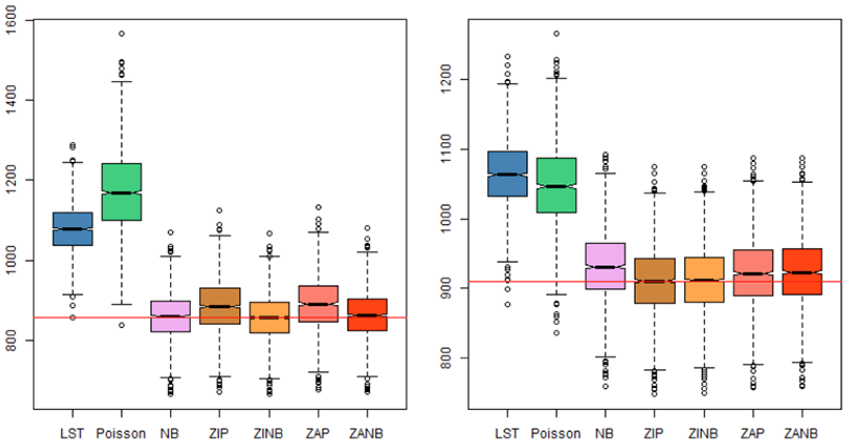


Figure 2.4. Boxplot of AIC from seven models (w = 0.6)

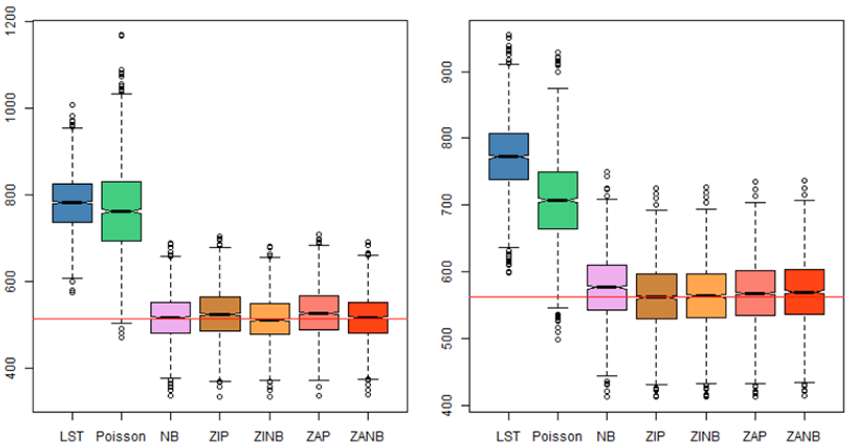


Figure 2.5. Boxplot of AIC from seven models (w = 0.8)

From the boxplots, we can see that when k = 1, the NB model and the ZINB model had much lower AIC values compared with the Poisson and the ZIP model. The difference in AIC values between zero-inflated models (i.e., ZIP and ZINB) and zero-altered models (i.e., ZAP and ZANB) showed a tendency to get smaller as there was an increase of zero-inflation and dispersion. AIC values for the ZINB model were always low across all conditions.

Results from empirical data analysis

Descriptive statistics such as means (and standard deviations) or frequencies (and percentages) for the variables of age, sex, UNHLTH and physical activity are presented in Table 3. Participants reported an average of 3.63 unhealthy days during the past 30 days with a variance of 36.84, which was much larger than the mean; and 44.67% of the participants reported 0 unhealthy days.

Table 3. Descriptive statistics for independent and dependent variables (n = 5670)

Variable		Mean	SD	Frequency (%)
Age (years)		55.03	16.87	
Sex	Male	2126		38.7
	Female	3362		61.3
# Unhealthy Days		3.63	6.07	
Physical Activity	Highly Active	1659		32.5
	Active	1059		20.8
	Insufficiently Active	1059		20.8
	Inactive	1323		25.9

Figure 3 presents the frequency plot of the response variable, UNHLTH. Notice that this variable showed an extremely right skewed distribution with a spike at zero.

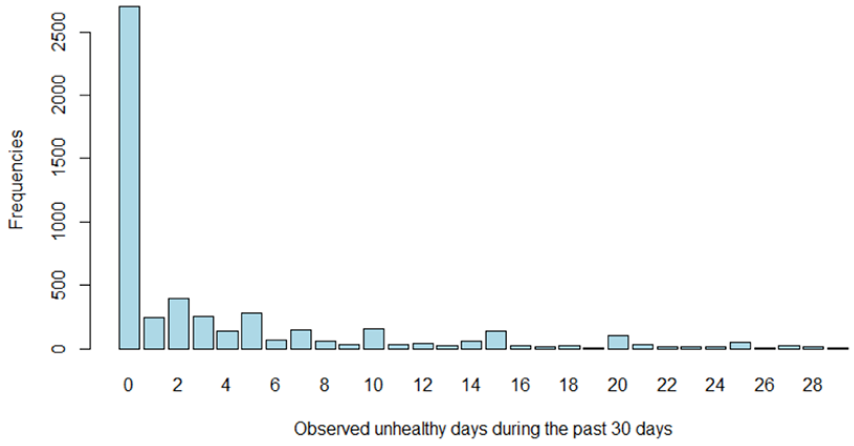


Figure 3. Frequency plot of the response variable UNHLTH from BRFSS data

Seven models described above were used to fit the data. AIC values and −2log-likelihood for each model are presented in Table 4.1. The Poisson regression model had the largest AIC values, demonstrating a poor fit to the data. Of the remaining six models, the NB, ZINB, and ZANB models had smaller AICs compared with the ZIP, ZAP, and LST models, indicating better fit with the data for the three negative binomial based models. ZINB and ZANB models yielded similar AICs and are considered as the best models even after penalizing the number of parameters in the model. Since not all of the models were nested with each other, under the null hypothesis that the models were indistinguishable, Vuong tests were used to further compare the above models. LST couldn't be compared because it has a different term for its dependent variable, i.e. it is log-transformed. The first comparison was made between the Poisson model and the NB model, with a Vuong test statistic of −42.41, and $p < 0.01$, indicating the NB model was more preferred. The more preferable model was then compared with the next model. After a series of tests and model comparisons (as shown in Table 4.2), ZANB was chosen as the best model. ZINB could be viewed as a second choice with a Vuong test statistic of −1.77, and $p = 0.04$ compared to ZANB, although the p-value was not within the range needed to control Type I error rate.

Table 4.1. Model fit comparison for the BRFSS data

	LST	Poisson	NB	ZIP	ZINB	ZAP	ZANB
AIC	24050.78	47932.45	21447.22	27814.26	21060.95	27814.26	21060.06
−2log-likelihood	24046.78	47908.45	21421.22	27766.26	21010.95	27766.26	21010.06
c	13	12	13	24	25	24	25

Note: AIC = the Akaike Information Criterion, and *c* is the number of free parameters in the model. LST = least squares regression with transformed outcome, Poisson = Poisson regression model, NB = negative binomial regression model, ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, ZAP = zero-altered Poisson model, ZANB = zero-altered negative binomial model.

Table 4.2. Vuong non-nested tests results for the BRFSS data

Model Comparison	Vuong Test Statistic	*p*	Preferable Model
Poisson vs. NB	−41.42	<0.01	NB
NB vs. ZIP	22.30	<0.01	NB
NB vs. ZINB	−12.16	<0.01	ZINB
ZINB vs. ZAP	25.35	<0.01	ZINB
ZINB vs. ZANB	−1.77	0.04	ZANB

Note: Poisson = Poisson regression model, NB = negative binomial regression model, ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, ZAP = zero-altered Poisson model, ZANB = zero-altered negative binomial model.

Table 5.1. Estimated regression coefficients (and standard errors) for LST, Poisson, and NB

Regressor	LST	*SE*	Poisson	*SE*	NB	*SE*
Intercept	0.713***	(0.040)	0.987***	(0.023)	0.983***	(0.080)
PA_active	0.032	(0.068)	0.097*	(0.038)	0.116	(0.134)
PA_insufficiently active	−0.004	(0.068)	0.021	(0.039)	0.027	(0.133)
PA_inactive	0.162**	(0.062)	0.360***	(0.033)	0.365**	(0.122)
SEX_female	0.117**	(0.053)	0.173***	(0.029)	0.178	(0.104)
AGE	−0.007***	(0.002)	−0.010***	0.000	−0.010**	(0.003)
PA_active*SEX_female	0.049	(0.086)	−0.002	(0.046)	−0.025	(0.169)
PA_insufficiently active*SEX_female	0.158	(0.085)	0.231***	(0.046)	0.225	(0.168)
PA_inactive*SEX_female	0.11	(0.080)	0.089*	(0.040)	0.083	(0.157)
PA_active*AGE	0.001	(0.003)	0.004**	(0.001)	0.005	(0.005)
PA_insufficiently active*AGE	0.005	(0.003)	0.009***	(0.001)	0.009	(0.005)
PA_inactive*AGE	0.007**	(0.002)	0.012***	(0.001)	0.012**	(0.005)

Note: "Male" was the reference group for sex and "highly active" was the reference group for physical activity. LST = least squares regression with transformed outcome, Poisson = Poisson regression model, NB = negative binomial regression model.

Table 5.2. Estimated regression coefficients (and standard errors) for ZIP, ZINB, ZAP, and ZANB under the Count Model

Regressor	ZIP	*SE*	ZINB	*SE*	ZAP	*SE*	ZANB	*SE*
Intercept	1.903***	(0.023)	1.754***	(0.065)	1.903***	(0.023)	1.753***	(0.065)
PA_active	0.047	(0.038)	0.051	(0.105)	0.047	(0.038)	0.055	(0.106)
PA_insufficiently active	0.000	(0.039)	−0.001	(0.106)	0.000	(0.039)	−0.001	(0.106)
PA_inactive	0.281***	(0.033)	0.325***	(0.095)	0.281***	(0.033)	0.326***	(0.095)
SEX_female	0.039	(0.030)	0.046	(0.082)	0.039	(0.030)	0.046	(0.082)
AGE	−0.002*	(0.001)	−0.002	(0.002)	−0.002*	(0.001)	−0.002	(0.002)
PA_active*SEX_female	−0.044	(0.047)	−0.047	(0.129)	−0.044	(0.046)	0.051	(0.129)
PA_insufficiently active*SEX_female	0.123**	(0.046)	0.143	(0.149)	0.123**	(0.046)	0.142	(0.129)
PA_inactive*SEX_female	0.015	(0.041)	0.007	(0.119)	0.015	(0.041)	0.005	(0.120)
PA_active*AGE	0.002	(0.001)	0.002	(0.004)	0.002	(0.001)	0.007	(0.003)
PA_insufficiently active*AGE	0.005***	(0.001)	0.005	(0.004)	0.005***	(0.001)	0.053	(0.004)
PA_inactive*AGE	0.006***	(0.001)	0.007*	(0.003)	0.006***	(0.001)	0.007*	(0.003)

Note: "Male" was the reference group for sex and "highly active" was the reference group for physical activity. For zero-inflated and zero-altered models, Count Model has relationship between covariates and count mean and Zero-inflation Model has relationship between covariates and probability of zeros. ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, ZAP = zero-altered Poisson model, ZANB = zero-altered negative binomial model. Significance levels: *** = 0.001, ** = 0.01, * = 0.05.

Table 5.3. Estimated regression coefficients (and standard errors) for ZIP, ZINB, ZAP, and ZANB under the Zero-Inflation Model

Regressor	ZIP	*SE*	ZINB	*SE*	ZAP	*SE*	ZANB	*SE*
Intercept	0.393***	(0.078)	0.127	(0.092)	−0.395***	(0.078)	−0.395***	(0.078)
PA_active	−0.074	(0.131)	−0.074	(0.151)	0.075	(0.131)	0.075	(0.131)
PA_insufficiently active	−0.018	(0.131)	−0.019	(0.150)	0.018	(0.130)	0.018	(0.130)
PA_inactive	−0.123	(0.120)	−0.060	(0.135)	0.125	(0.120)	0.125	(0.120)
SEX_female	−0.126*	(0.102)	−0.256*	(0.118)	0.236*	(0.102)	0.236*	(0.102)
AGE	0.015***	(0.003)	0.017***	(0.004)	−0.015***	(0.003)	−0.015***	(0.003)
PA_active*SEX_female	−0.103	(0.165)	−0.129	(0.193)	0.102	(0.165)	0.102	(0.165)
PA_insufficiently active*SEX_female	−0.226	(0.164)	−0.223	(0.192)	0.228	(0.164)	0.228	(0.164)
PA_inactive*SEX_female	−0.170	(0.154)	−0.184	(0.175)	0.170	(0.154)	0.170	(0.154)
PA_active*AGE	−0.002	(0.005)	−0.001	(0.006)	0.002	(0.005)	0.002	(0.005)
PA_insufficiently active*AGE	−0.008	(0.005)	−0.007	(0.006)	0.008	(0.005)	0.008	(0.005)
PA_inactive*AGE	−0.010*	(0.004)	−0.010*	(0.005)	0.010*	(0.004)	0.010*	(0.004)

Note: "Male" was the reference group for sex and "highly active" was the reference group for physical activity. For zero-inflated and zero-altered models, Count Model has relationship between covariates and count mean and Zero-inflation Model has relationship between covariates and probability of zeros. ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, ZAP = zero-altered Poisson model, ZANB = zero-altered negative binomial model. Significance levels: *** = 0.001, ** = 0.01, * = 0.05.

Regression coefficients and standard errors were estimated and presented in Table 5.1 and 5.2 for each of the seven models when applied to the BRFSS dataset. Standard errors estimated from different models were quite different. There was a tendency for the worse models to have smaller standard errors. For instance, although estimates from the Poisson model were similar to those from the NB model, their standard errors were much smaller, thus yielding significant results for most of the regressors, which was most likely not accurate. It was the same when comparing ZIP versus ZINB and ZAP versus ZANB.

With PA (i.e., physical activity), gender, age, PA*gender, and PA*age predicting both the count model and zero-inflation model, Table 5.2 shows parameter estimates from the ZANB model (the final model). Participants in the highly active group and males were used as reference groups. After controlling for age, gender, and their interaction terms with PA, compared with highly active people, inactive people were likely to experience 1.39 (= exp(0.326), $p < 0.001$) more unhealthy days. This trend can also be seen in Figure 4, where both inactive males and females had higher means of UNHLTH than other groups of participants. (Male) gender (odds ratio = 1.27, $p < 0.05$) and younger age (odds ratio = 0.99, $p < 0.001$) were the only results to be significant predictors for those who experienced 0 unhealthy days versus those who experienced more than 0 unhealthy days. Thus, females and older people were more likely to report

unhealthy days, although it should be pointed out that the odds ratio for age was not very meaningful in size, even if significant.

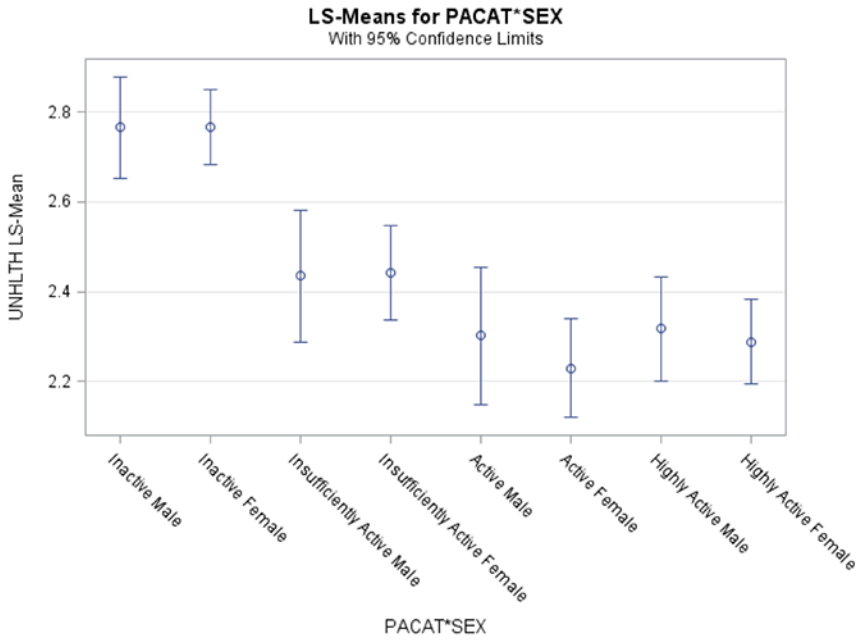


Figure 4. Least-squared Means of UNHLTH by PA and Gender with 95% Confidence Limits

Discussion

This study evaluated seven regression models under various conditions of zero-inflation and dispersion by analyzing simulated datasets and an empirical dataset. Results from both studies suggested that when the data include excessive zeros (even as low as 20%) and over-dispersion, zero inflated models (i.e. ZIP, ZINB, ZAP, and ZANB) perform better than Poisson regression and ordinary least-squares regression with transformed outcomes (LST). It was only when fitting data with no zero-inflation and the least dispersion (i.e., $w = 0.0$ and $k = 50$) in the simulation study, that the Poisson regression model performed well and had the highest selection rate.

The poor fit from the LST might be that the log-transformation still fails to correct the non-normality and to address the inflation of zeros. Another drawback

of using a transformation is that the regression coefficients are harder to interpret. The Poisson distribution is the probability model usually assumed for count data, however, zero-inflated count data usually tend to have much bigger variance than the mean, which violates its assumption that the mean equals the variance. In both cases, when failing to address the problem of zero-inflation and over-dispersion, standard errors of the estimates tended to be deflated or under estimated (Hilbe, 2014). Furthermore, if inappropriately choosing the LST or the Poisson model, there is greater tendency to make Type I errors, i.e. a variable may appear to be a significant predictor when it is in fact, not significant. Estimated regression coefficients from Table 5.1 demonstrate this kind of bias.

Results from these studies of simulated and real data support using special zero inflated models for zero-inflated data. When over-dispersion also exists even in the non-zero part of the data, a negative binomial regression instead of the regular Poisson regression should be used. Compared with other models, the ZINB model had the most consistent performance at any combination of dispersion and zero-inflation in the simulation study. The use of zero inflated models can be justified on both substantive and statistical grounds. Substantively, zero inflated models have the ability to identify the factors that have significant effects on the probability that the participant is from the non-susceptible group by means of a binary regression model; and the magnitude of the counts given that the participant is from the susceptible group by means of a Poisson regression or negative binomial regression. Factors or explanatory variables do not need to be the same for the binomial model and the count model. Although the NB model can also effectively offer accurate estimation under some degrees of zero-inflation and over-dispersion, it cannot provide information about possible mechanisms underlying the zero-inflation. Statistically, zero inflated models provide more accurate estimates as shown by both the simulation results and empirical data analysis results.

Zero-inflated models are more preferred than zero-altered models when we assume zeros can be produced both from the zero-inflation process and the count process. In the simulation study, data were generated under this mechanism and we found that zero-inflated models out-performed zero-altered models, especially when the levels of zero-inflation and dispersion were low. Therefore, the decision when choosing between these two should rely on the nature of the research questions. The biggest difference between them is that zero-inflated models distinguish between structural zeros (true zeros) and random zeros (false zeros), although zero-altered models do not. In public health and medicine studies, zero-inflated models may be conceptualized as allowing zeros to arise from at-risk

(susceptible) and not-at-risk (non-susceptible) populations. In contrast, we may conceptualize zero-altered models as having zeros only from an at-risk population (Rose et al., 2006). For instance, when answering a survey question that asks the number of drinks someone had during the past month, some people report 0 drinks because they are abstainers and they never drink. However, for people who are regular drinkers, they might also report 0 drinks if they did not drink during that month. As mentioned earlier, these latter zero responses are called random zeros (or false zeros) (Zuur, et al., 2009). It is more appropriate to use ZIP and ZINB in these kind of situations when the study design has a greater chance of having random zeros.

Another interest of the study with empirical data was to explore the relationship between health related quality of life (HRQoL) and physical activity (PA). Many research studies have shown that PA helps to improve overall health and fitness, and reduce risk of health conditions including diabetes, coronary heart disease, stroke, and cancers (CDC, 2014). Despite the well-known benefits of exercise, according to the CDC, less than half of American adults meet the recommended level of PA. HRQoL describes both the physical and mental well-being of an individual. It is an important concept in health research and can help to inform decisions on the prevention and treatment of diseases. The present study examined the relationship between PA and HRQoL after controlling for relevant demographic characteristics within the context of a large representative health survey from Rhode Island. Results showed that participants reporting higher levels of PA tended to report fewer unhealthy days. Specifically, compared with participants in the highly active group, those who seldom reported any physical activity were likely to experience 1.30 more unhealthy days. Females and older people were also more likely to report unhealthy days versus 0 unhealthy day compared to males and younger people. These findings offer a better understanding that health-related lifestyle behaviors, such as being more physically active, can improve HRQoL and might help to inform policy makers to provide more intervention programs for the general population.

There were also some limitations of the study. First, for the empirical study, explanatory variables for the zero versus non-zero model and the count model were set to be the same. The most attractive advantage of using zero-inflated models is that they allow researchers to have different predictors for two parts of the models, which usually can be justified theoretically. Second, since the data were collected via a telephone survey, various response biases and non-response biases would occur. For instance, participants consisted mostly of older people with an average age of 55.51 years; thus, the sample was not sufficiently random.

Third, the cross-sectional nature of the data was another limitation of the study. Since these data were cross-sectional, no temporal order can be determined, so it is possible that those with higher health-related quality of life (HRQoL) reported more physical activity (PA). Future longitudinal designs are needed to tease out temporal relationships. Only age and gender were controlled for in the empirical data analysis. It is possible that other unmeasured factors, such as disease states and seasonality, could be potential confounding variables of the relationship between PA and HRQoL. Future longitudinal analyses would help to improve our understanding of these relationships and increase the predictive power of the study, in addition to what model is used to examine the data. Finally, the UNHLTH ranges from 0 to 29 days, which follows a zero-inflated negative binomial distribution truncated at 29. Creel and Loomis (1990) suggest that accounting for truncation of the response variable provides a more accurate coefficient estimates, regardless of the choice of the statistical model. Although a truncated model was not used in this study, it might be of interest in future studies.

Acknowledgements

This research was supported in part by G20RR030883 from the National Institutes of Health.

References

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2012). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors, 27*(1), 166-177. doi: 10.1037/a0029508.

Beydoun, M. A., Beydoun, H. A., Boueiz, A., Shroff, M. R., & Zonderman, A. B. (2012). Antioxidant status and its association with elevated depressive symptoms among US adults: National Health and Nutrition Examination Surveys 2005-6. *British Journal of Nutrition, 109*(09), 1714-1729. doi: 10.1017/S0007114512003467.

Bolker, B. M. (2008). *Ecological models and data in R.* Princeton, NJ: Princeton University Press.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44*(1), 62-91. doi: 10.1006/jmps.1999.1277

Centers for Disease Control and Prevention (CDC). (2012). *Behavioral risk factor surveillance system*. Retrieved from http://www.cdc.gov/brfss/

Centers for Disease Control and Prevention (CDC). (2014). *Facts about physical activity*. Retrieved from http://www.cdc.gov/physicalactivity/data/facts.html

Creel, M. D. & Loomis, J. B. (1990). Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. *American Journal of Agricultural Economics, 72*(2), 434-441. doi: 10.2307/1242345

Famoye, F. & Singh, K. P. (2006). Zero-inflated generalized Poisson model with an application to domestic violence data. *Journal of Data Science, 4*(1), 117-130. doi: 10.1177/1471082X0700700202

Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. *Working Paper EC-94-10*. Leonard N. Stern School of Business, New York University.

Hilbe, J. M. (2014). *Mode g count data*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139236065

Jackman, S. (2008). *PSCL: classes and methods for R developed in the political science computational laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California. R package version 1.04.4. Available at http://CRAN.R-project.org/package=pscl.

Karazsia, B. T. & van Dulmen, M. H. (2008) Regression models for count data: illustrations using longitudinal predictors of childhood injury. *Journal of Pediatric Psychology, 33*(10), 1076-1084. doi: 10.1093/jpepsy/jsn055

King, G. (1989). Event count models for international relations: generalizations and applications. *International Studies Quarterly, 33*(2), 123-147. doi: 10.2307/2600534

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing, *Technometrics, 34,* 1–14. doi: 10.2307/1269547

Lin, T. H. & Tsai, M. H. (2013). Modeling health survey data with excessive zero and *K* responses. *Statistics in Medicine, 32*(9), 1572-1583. doi: 10.1002/sim.5650.

Liu, H. & Power, D. A. (2007). Growth curve models for zero-inflated count data: an application to smoking behavior. *Structural Equation Modeling, 14*, 247–79. doi: 10.1080/10705510709336746.

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, *65*(1), 163-180. doi: 10.1111/j.2044-8317.2011.02031.x.

Long, J. (1997). *Regression models for categorical and limited dependent variables*. CA: Thousand Oaks, Sage.

Mahalik, J. R., Levine Coley, R., McPherran Lombardi, C., Doyle Lynch, A., Markowitz, A. J., & Jaffee, S. R. (2013). Changes in health risk behaviors for males and females from early adolescence through early adulthood. *Health Psychology, 32*(6), 685-694. doi: 10.1037/a0031658.

Mullahy, J. (1986). Specifications and testing of some modified count data model. *Journal of Econometrics, 33*(3), 341–365. doi: 10.1016/0304-4076(86)90002-3.

Min, Y. & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling, 5,* 1–19. doi: 10.1191/1471082X05st084oa.

Ma, R., Hasan, M. T., & Sneddon G. (2009). Modeling heterogeneity in clustered count data with extra zeros using compound Poisson random effect. *Statistics in Medicine, 28*(18), 2356–2369. doi: 10.1002/sim.3619.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Roebuck, M. C., French, M. T., & Dennis, M. L. (2004). Adolescent marijuana use and school attendance. *Economics of Education Review, 23*(2), 133–141. doi: 10.1016/s0272-7757(03)00079-7

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics, 16*(4), 463-481. doi: 10.1080/10543400600719384.

Schunck, R. & Rogge, B. G. (2012). No causal effect of unemployment on smoking? A German panel study. *International Journal of Public Health, 57*(6), 867-874. doi: 10.1007/s00038-012-0406-5.

Vives, J., Losilla, J. M., & Rodrigo, M. F. (2006). Count data in psychological applied research. *Psychological Reports, 98*(3), 821–835. doi: 10.2466/PR0.98.3.821-835.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, *57*(2), 307-333. doi: 10.2307/1912557

Warton, D. I. (2005). Many zeros does not mean zero-inflation: Comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, *16*, 275-289. doi: 10.1002/env.702.

Williamson, J. M., Lin, H., Lyles, R. H., & Hightower, A. W. (2007). Power calculations for ZIP and ZINB models. *Journal of Data Science*, *5*(4), 519-534.

Wong, K. Y. & Lam, K. F. (2012). Modeling zero-inflated count data using a covariate-dependent random effect model. *Statistics in Medicine*, *32*(8), 1283-1293. doi: 10.1002/sim.5626

Zorn, C. (1996). Evaluating zero-inflated and Hurdle Poisson specifications, *Midwest Political Science Association, 18*(20), 1–16.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. NY: Springer.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 544.

ISSN 1538 – 9472

Emerging Scholars

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 545-577.
doi: 10.22237/jmasm/1493598000

ISSN 1538 − 9472

Graphical Log-Linear Models: Fundamental Concepts and Applications

Niharika Gauraha
Indian Statistical Institute, Bangalore Center
Bangalore, India

A comprehensive study of graphical log-linear models for contingency tables is presented. High-dimensional contingency tables arise in many areas. Analysis of contingency tables involving several factors or categorical variables is very hard. To determine interactions among various factors, graphical and decomposable log-linear models are preferred. Connections between the conditional independence in probability and graphs are explored, followed with illustrations to describe how graphical log-linear model are useful to interpret the conditional independences between factors. The problem of estimation and model selection in decomposable models is discussed.

Keywords: Graphical log-linear models, contingency tables, decomposable models, hierarchical log-linear models

Introduction

The aim in the current study is to provide insight into graphical log-linear models (LLMs) by providing a concise explanation of the underlying mathematics and statistics, by pointing out relationships to conditional independence in probability and graphs, and providing pointers to available software and important references. LLMs are the most widely used models for analyzing cross-classified categorical data (Christensen, 1997). LLM supports various ranges of models based on non-interaction assumptions. For fairly large-dimensional tables, the analysis becomes difficult; as the number of factors increases the number of interaction terms grows exponentially. Graphical LLMs are a way of representing relationships among the factors of a contingency table using a graph. The graphical LLMs have two great advantages: from the graph structure, it is easy to read off the conditional independence relations; and graph-based algorithms usually provide efficient computational algorithms for parameter estimation and model selection.

Niharika Gauraha is a PhD student. Email them at: niharika.gauraha@gmail.com.

The decomposable LLMs are a restricted class of GLLMs which are based on chordal graphs. There are several reasons for using decomposable models over an ordinary GLLM. Firstly, the maximum likelihood estimates can be found explicitly. Secondly, closed-form expressions exist for the test statistics. Another advantage is that it has triangulated graph-based efficient inference algorithms. Thus decomposable models are mostly used for analysis of high-dimensional tables.

Graph Theory and Markov Networks

Graph Theory

Necessary concepts of graph theory that will be used are discussed. See West (2000) for further details on graph theory. A graph G is a pair $G = (V, E)$, where V is a set of vertices and E is a set of edges. A graph is said to be an undirected graph when E is a set of unordered pairs of vertices. Consider only a simple graph that has neither loops nor multiple edges.

Definition 1 (Boundary): Let $G = (V, E)$ be an undirected graph. The neighbors or boundary of a subset A of vertices is a subset C of vertices such that all nodes in C are not in A but are adjacent to some vertex in A.

$$\mathrm{bd}(\mathrm{A}) = \{u \in \mathrm{V} \setminus \mathrm{A} | \ \exists v \in \mathrm{A} : \{u, v\} \in \mathrm{E}\}$$

Definition 2 (Maximal Clique): A clique of a graph G is a subset C of vertices such that all vertices in C are mutually adjacent. A clique is said to be maximal if no vertex can be added to C without violating the clique property.

Definition 3 (Chordal (Triangulated) Graphs): In graph theory, a chord of a cycle C is defined as an edge which is not in the edge set of C but joins two vertices from the vertex set C. A graph is said to be a chordal graph if every cycle of length four or more has a chord.

Definition 4 (Isomorphic Graphs): Two graphs are said to be isomorphic if they have same number of vertices, same number of edges, and they are connected in the same way.

Conditional Independence

The concept of conditional independence in probability theory is very important and it is the basis for the graphical models. It is defined as follows:

Definition 5 (Conditional Independence): Let X, Y, and Z be random variables with a joint distribution P. The random variables X and Y are said to be conditionally independent given the random variable Z if and only if the following holds:

$$\mathrm{P}(X, Y \mid Z) = \mathrm{P}(X \mid Z)\mathrm{P}(Y \mid Z)$$
$$\mathrm{P}(X \mid YZ) = \mathrm{P}(X \mid Z)$$

Dawid's (1979) notation, $X \perp\!\!\!\perp Y \mid Z$, is also used. Conditional independence has a vast literature in the field of probability and statistics; see also Pearl and Paz (1987).

Markov Networks and Markov Properties

Markov network graphs, Markov networks, and different Markov properties for the Markov Networks are now defined.

Definition 6 (Markov Network Graphs): A Markov network graph is an undirected graph $G = (\mathrm{V}, \mathrm{E})$ where $\mathrm{V} = \{X_1, \ldots, X_n\}$ represents random variables of a multivariate distribution.

Definition 7 (Markov Networks): A Markov network M is a pair $M = (G, \Psi)$. Where G is a Markov network graph and $\Psi = \{\psi_1, \ldots, \psi_m\}$ is a set of non-negative functions for each maximal clique $C_i \in G \; \forall i = 1, \ldots, m$, and the joint probability density function (pdf) can be decomposed into factors as

$$\mathrm{P}(x) = \frac{1}{Z} \prod_{a \in C_m} \psi_a(x)$$

where Z is a normalizing constant.

Definition 8 (Pairwise Markov Property (P)): A probability distribution P satisfies the pairwise Markov property for a given undirected graph G if, for every pair of non-adjacent vertices X and Y, X is independent of Y given the rest.

$$X \perp\!\!\!\perp Y \mid (V \setminus X, \mathrm{Y})$$

Definition 9 (Local Markov Property (L)): A probability distribution P satisfies the local Markov property for a given undirected graph G if every variable X is conditionally independent of its non-neighbors in the graph, given its neighbors.

$$X \perp\!\!\!\perp (V \setminus (X \cup \mathrm{bd}(X))) \mid \mathrm{bd}(X)$$

Definition 10 (Global Markov Property (G)): A probability distribution P is said to be global Markov with respect to an undirected graph G if and only if, for any disjoint subsets of nodes A, B, and C such that C separates A and B on the graph, the distribution satisfies the following:

$$\mathrm{A} \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C}$$

Note the above three Markov properties are not equivalent to each other. The local Markov property is stronger than the pairwise one, while weaker than the global one. More precisely,

Proposition 1: For any probability measure the following holds:

$$(\mathrm{G}) \Rightarrow (\mathrm{L}) \Rightarrow (\mathrm{P})$$

See Lauritzen (1996), for proof of Proposition 1. Refer to Lauritzen (1996) and Edwards (2000) for further details on graphical models, and to Darroch, Lauritzen, and Speed (1980) for details on Markov fields for LLMs.

Notations and Assumptions

The notations and the assumptions are now discussed. Consider three-dimensional tables for notational simplicity; this is also a true representative of k-dimensions and thus can be easily extended to any higher dimensions by increasing the

number of subscripts. See Christensen (1977) and Bishop, Fienberg, and Holland (1989).

Consider a three-dimensional table with factors X, Y, and Z. Numeric $\{1, 2, 3\}$ and alphabetic $\{X, Y, Z\}$ symbols are used interchangeably to represent the factors of a contingency table. Suppose the factors X, Y, and Z have I, J, and K levels, respectively. Then we have an $I \times J \times K$ contingency table.

The following notations are defined for each elementary cell (i, j, k) for $i = 1,\ldots, I$, $j = 1,\ldots, J$, and $k = 1,\ldots, K$:

n_{ijk} = the observed counts in the cell (i, j, k)
m_{ijk} = the expected counts in the cell (i, j, k)
$\hat{m}_{ijk}$ = the Maximum Likelihood Estimate (MLE) of m_{ijk}
p_{ijk} = the probability of a count falling in cell (i, j, k)
$\hat{p}_{ijk}$ = the MLE of p_{ijk}

The following notations are used for sums of elementary cell counts, where "." represents summation over that factor. For example,

$$n_{i..} = \sum_{jk} n_{ijk}$$
$$n_{i.k} = \sum_{j} n_{ijk}$$
$$N = n_{...} = \text{total number of observations}$$

Similarly, the marginal totals of probabilities and the expected counts are denoted by $p_{.jk}$, and $m_{.jk}$, etc.

Denote by C the tables of sums obtained by summing over one or more factors, e.g. C_{12} represents tables of counts $n_{ij.}$. Subscripted u-term notation is used for main effects and interactions. For example, u_{ij} is used for two-factor interactions $\forall i = 1,\ldots, I$ and $\forall j = 1,\ldots, J$. We may interchangeably use $u_{12(ij)}$ and u_{ij}; the latter is obtained by simply dropping the second set of subscript. Thus

$$u_{12} = u_{12(ij)} \ \forall i = 1,\ldots, I, j = 1,\ldots, J$$

Assume that the observed cell counts are strictly positive for all models we consider throughout this article.

Overview of Contingency Tables

A contingency table is a table of counts that summarizes the relationship between factors. In a multivariate qualitative data set where each individual is described by a set of attributes, all individual with same attributes are counted; this count is entered into a cell of a corresponding contingency table (see Bishop, Fienberg, & Holland, 1989). The term contingency was introduced by Pearson (1904). There is an extensive body of literature on contingency tables; see A. H. Andersen (1974), Bartlett (1935), and Goodman (1969).

Example 1: Table 1 provides an example of a three-dimensional contingency table taken from example 3.2.1 of Christensen (1997).

Types of Contingency Tables

Based on the underlying assumption of sampling distributions, contingency tables are divided into three main categories as follows:

The Poisson Model In this model, it is assumed that cell counts are independent and Poisson-distributed. The total number of counts and the marginal counts are random variables. For three-dimensional tables with counts as random variables, the joint probability density function (pdf) can be written as

$$\mathrm{f}\left(\{n_{ijk}\}\right) = \prod_i \prod_j \prod_k \frac{m_{ijk}^{n_{ijk}} \mathrm{e}^{-m_{ijk}}}{m_{ijk}!} \tag{1}$$

The Multinomial Model In this model, it is assumed that the total number of subjects N is fixed. With this constraint imposed on independent Poisson distributions, the cell counts yield a multinomial distribution. For proof we refer to Fisher (1922). The pdf for this model is given as

Table 1. Personality type table

		Diastolic Blood Pressure	
Personality Type	**Cholesterol**	**Normal**	**High**
A	Normal	716	79
	High	207	25
B	Normal	819	67
	High	186	22

$$\mathrm{f}\left(\left\{n_{ijk}\right\}\right)=\frac{N!}{\prod_i\prod_j\prod_k n_{ijk}!}\prod_i\prod_j\prod_k\left(\frac{m_{ijk}}{N}\right)^{n_{ijk}} \tag{2}$$

The Product-Multinomial Model In this model, it is assumed that one set of marginal counts is fixed and the corresponding table of sums follow a product-multinomial distribution. For example, consider a three-dimensional table with total counts for the first factor, $n_{.jk}$, fixed. The pdf is given as

$$\mathrm{f}\left(\left\{n_{ijk}\right\}\right)=\prod_j\prod_k\left[\frac{n_{.jk}!}{\prod_i n_{ijk}!}\prod_i\left(\frac{m_{ijk}}{n_{ijk}}\right)^{n_{ijk}}\right] \tag{3}$$

Introduction to Log-Linear Models

As discussed previously, the distribution of cell probabilities belong to exponential family (Poisson, multinomial, and product-multinomial). Construct a linear model in the log scale of the expected cell count. A LLM for a three-factor table is defined as

$$\log\left(m_{ijk}\right)=u+u_{1(i)}+u_{2(j)}+u_{3(k)}+u_{12(ij)}+u_{13(ik)}+u_{23(jk)}+u_{123(ijk)} \tag{4}$$

with the following identifiability constraints:

$$\begin{aligned}
&\sum_i u_{1(i)}=\sum_j u_{2(j)}=\sum_k u_{3(k)}=0\\
&\sum_i u_{12(ij)}=\sum_j u_{12(ij)}=0\\
&\sum_j u_{12(ik)}=\sum_k u_{12(ik)}=0\\
&\sum_j u_{12(jk)}=\sum_k u_{12(jk)}=0\\
&\sum_i u_{123(ijk)}=\sum_j u_{123(ijk)}=\sum_k u_{123(ijk)}=0
\end{aligned}$$

The above model is called saturated or unrestricted because it contains all possible one-way, two-way, and three-way effects. In general, if no interaction terms are set to zero, it is called the saturated model.

The number of terms in a LLM model depends on the dimensions or number of factors and the interdependencies between the factors; it does not depend on the number of cells (see Birch, 1963 for more details). The model given by equation (4) applies to all three kinds of contingency tables with three factors (as discussed in the previous section), but there may be differences in the interpretations of the interaction terms (see Kreiner, 1998; Lang, 1996b). There is a wide body of literature on LLMs, see for instance Agresti (2002), Christensen (1997), Zelterman (2006), and Knoke and Burke (1980).

Log-Linear Models as Generalized Linear Models

Recall the generalized linear model (GLM). It consists of a linear predictor and a link function. The link function determines the relationship between the mean and the linear predictor. Here, we show that the LLMs are special instances of GLMs for Poisson-distributed data; see Nelder and Wedderburn (1972) for details.

Consider a 2 × 2 Poisson model with two factors, say X and Y, and suppose cell counts n_{ij} are response variables such that $n_{ij} \sim \text{Poisson}(m_{ij})$ and the factors X and Y are explanatory variables. Define a link function g as $g(m_{ij}) = \log(m_{ij})$. The linear predictor is defined as $\mathbf{X}'\boldsymbol{\beta}$, where $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ is the vector of unknown parameters. For this model, $\mathbf{X}$ and $\boldsymbol{\beta}$ are defined as

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \end{bmatrix}$$

The model can be expressed as follows:

$$\log\left(m_{ij}\right) = x_i'\boldsymbol{\beta} = \mu + \alpha_i + \beta_j + \left(\alpha\beta\right)_{ij}$$

Rename the parameters as

$$\log\left(m_{ij}\right) = u + u_1 + u_2 + u_{12}$$

The above model is the same as the LLM defined for two-factor tables, where u is the overall mean, u_1 and u_2 are the main effects, and u_{12} is the interaction effect. LLMs can be fit as generalized linear models by using software packages available for GLMs, e.g. the glm() function in the stats R package.

Classes of Log-Linear Models

Comprehensive Log-Linear Models

The class of comprehensive LLMs is defined as follows:

Definition 11 (Comprehensive Log-Linear Models): A log-linear model is said to be comprehensive if it contains the main effects of all the factors.

For example, a comprehensive LLM for the three-factor contingency tables must include all the main effects u_1, u_2, and u_3, along with other interaction effects, if any (see Zelterman, 2006).

Hierarchical Log-Linear Models

The class of hierarchical LLMs is defined as follows:

Definition 12 (Hierarchical Log-Linear Models): A LLM is said to be hierarchical if it contains all the lower-order terms which can be derived from the variables contained in a higher-order term.

For example, if a model for three-dimension table includes u_{12}, then u_1 and u_2 must be present. Conversely, if $u_2 = 0$, then we must have $u_{12} = u_{23} = u_{123} = 0$. The hierarchical models may be represented by giving only the terms of highest order, also known as a generating class, because all the lower-order terms are implicit. The generating class is defined as follows:

Definition 13 (Generating class): The highest-order terms in hierarchical LLMs are called a generating class because they generate all of the lower-order terms in the model.

Example 2: A LLM with generating classes C = {[123], [34]} corresponds to the following log-linear model:

$\log(m_{hijk}) = u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{23} + u_{13} + u_{123} + u_{34}$
Members of generating class [123] = {[1], [2], [3], [12], [23], [13], [123]}
Members of generating class [34] = {[3], [4], [34]}

All models considered in the remaining sections of this article are hierarchical and comprehensive LLMs unless stated otherwise.

Graphical Log-Linear Models

Consider a class of LLMs that can be represented by graphs, called graphical log-linear models (GLLMs).

Definition 14 (Graphical Log-Linear Models): A LLM is said to be graphical if it contains all the lower-order terms which can be derived from variables contained in a higher-order term, the model also contains the higher order interaction.

For example, if a model includes u_{12}, u_{23}, and u_{31}, then it also contains the term u_{123}. In GLLMs, the vertices correspond to the factors and the edges correspond to the two-factor interactions. But the factors (vertices) and the two-factor interactions (edges) alone do not specify the graphical models. As mentioned previously, factorization of the probability distribution with respect to a graph must satisfy the Markov properties. For such a graph that respects the Markov properties with respect to a probability distribution, there is a one-to-one correspondence between GLLMs and graphs. It follows that every GLLM determines a graph and every graph determines a GLLM, as is illustrated by the following examples:

Example 3: Consider the model [123] [134]. The two-factor terms generated by [123] are [12], [13], and [23]. Similarly, the two-factor terms generated by [134] are [13], [14], and [34]. The corresponding graph is as given in Figure 1.

Conversely, read the LLM directly from the corresponding graph. Consider a graph as given in Figure 2; the edges are [12], [23], [13], and [34]. Because the generating class for the terms [12], [23], and [13] is the term [123], we must include [123] in the model. Hence, the corresponding GLLM is [123] [34].

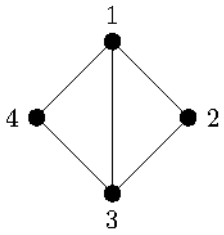


Figure 1. Graphical model of [123] [134]

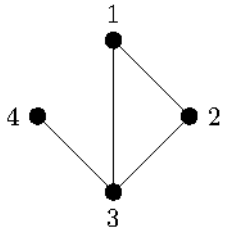


Figure 2. Graphical model of [123] [34]

Generating classes of GLLMs are in a one-to-one correspondence with the maximal cliques of the corresponding graph. Not all hierarchical LLMs have graphical representation. For example, the model [12] [13] [23] is hierarchical but it is not graphical because it does not contain the higher order term [123].

Decomposable Models Consider the class of decomposable models, which is a subclass of the GLLMs.

Definition 15 (Decomposable Log-Linear Models): A LLM model is decomposable if it is both graphical and chordal.

The main advantage of this model over other models is that it has closed form Maximum Likelihood Estimates (MLEs). For example, consider a decomposable model as given by Figure 1. The only conditional independence implied by the graph is that, given the factors 1 and 3, factors 2 and 4 are independent. The MLEs for the expected cell counts are factorized in a closed form in the terms of sufficient statistics as

$$\hat{m}_{ijkl} = \frac{n_{hij.} n_{h.jk}}{n_{h.j.}}$$

The derivation of MLE expressions, like the one above, is discussed in detail in a later section. For all the possible non-isomorphic graphical and decomposable models for the four-factor contingency tables, see Table 18 in the Appendix.

A few important articles concerned with the decomposable models are Goodman (1970, 1971b), Haberman (1974), Lauritzen, Speed, and Vijayan (1984), Meeden, Geyer, Lang, and Funo (1998) and Dahinden, Kalisch, and Bühlmann (2010).

Statistical Properties of the Log-Linear Models

Consider statistical properties of the hierarchical LLMs, like the existence of sufficient statistics, uniqueness of the MLE, and model testing.

The Sufficient Statistics for LLMs

The sufficient statistics exist for the hierarchical LLMs and are very easy to obtain. Consider the saturated model with simple multinomial sampling distribution for the three-factor contingency tables. The log-likelihood function of the multinomial is obtained from the pdf given by equation (1) as follows:

$$\log\left(\mathrm{f}\left(\left\{n_{ijk}\right\}\right)\right)=\log\left(\frac{N!}{\prod_i\prod_j\prod_k n_{ijk}}\right)+\sum_i\sum_j\sum_k n_{ijk}\log\left(m_{ijk}\right)-N\log N \quad (5)$$

Or, equivalently,

$$\log\left(\mathrm{f}\left(\left\{n_{ijk}\right\}\right)\right)=\sum_i\sum_j\sum_k n_{ijk}\log\left(m_{ijk}\right)+C \quad (6)$$

where C represents the constant terms. Substituting the value for $\log(m_{ijk})$ as given by equation (4),

$$\log\left(\mathrm{f}\left(\left\{n_{ijk}\right\}\right)\right)=\sum_i\sum_j\sum_k n_{ijk}\left(u+u_1+u_2+u_3+u_{12}+u_{13}+u_{23}+u_{123}\right)+C$$

The above expression can be also written as

$$\mathrm{f}\left(\left\{n_{ijk}\right\}\right)=\exp\left(Nu+\sum_i u_1 n_{i..}+\sum_j u_2 n_{.j.}+\sum_k u_3 n_{..k}+\sum_i\sum_j u_{12}n_{ij.}+\sum_i\sum_k u_{13}n_{i.k}\right.$$
$$\left.+\sum_j\sum_k u_{23}n_{.jk}+\sum_i\sum_j\sum_k u_{123}n_{ijk}+C\right)$$

Because the multinomial distribution belongs to exponential family sufficient statistic exists, see E. B. Andersen (1970). From the above expression it is apparent that, for the three-factor saturated model, the full table itself is the sufficient statistic since the lower-order terms are redundant and it will be

subsumed in the full table. The marginal sub-tables which correspond to the set of generating classes are the sufficient statistics for the log-linear models (see Birch, 1963).

Example 4: Consider a four-factor table with the following generating classes:

$$\{C_1, C_2\} = \{[123], [34]\}$$

Then $C_1(n) = [n_{ijk.}]$ is a three-dimensional marginal sub-table and $C_2(n) = [n_{..kl}]$ is a two-dimensional marginal sub-table. These two marginal sub-tables are the sufficient statistics for this model. For more details and proofs on the sufficient statistics for hierarchical LLMs, see Haberman (1973).

Maximum Likelihood Estimates for the LLMs

A unique set of MLEs for every cell count can be obtained from the sufficient statistics alone; see Birch (1963) for the proof. The Birch criteria are:

1. The marginal sub-tables obtained by summing over the factors not present in the max-cliques are the sufficient statistics for the corresponding expected cell counts. e.g., for the model [123] [34], $C_1(n) = [n_{ijk.}]$ and $C_2(n) = [n_{..kl}]$ are sufficient statistics for $m_{ijk.}$ and $m_{..kl}$, respectively.
2. All the sufficient statistics must be the same as the corresponding marginal sub-tables of their estimate means.

   $$C_i(\hat{m}) = C_i(n)$$

   for all i from 1 to the number of generating classes. e.g., for the model [123] [34], the estimated cell counts are

   $$\hat{m}_{ijk.} = n_{ijk.}$$
   $$\hat{m}_{..kl} = n_{..kl}$$

   Finally, the MLE of the expected cell counts for the model [123] [34] is expressed as follows:

$$\hat{m}_{ijkl} = \frac{n_{ijk.} n_{..kl}}{n_{..k.}}$$

The closed form expressions for the MLEs will be derived below in terms of sufficient statistics for three-factor contingency tables.

The reason for choosing MLE for computing the expected cell counts is its consistency and efficiency in large samples. There is extensive research on the MLEs of LLMs; see for example Glonek, Darroch, and Speed (1988), A. H. Andersen (1974), Haberman (1974), Meeden, Geyer, Lang, and Funo (1998), Birch (1963), Fienberg and Rinaldo (2007), Lang (1996a), Lang, McDonald, and Smith (1999), and Darroch (1962).

Testing Models

The assessment of a model's fit is very important as it describes how well it fits the data. Consider the following test statistics:

Pearson's χ^2 Statistic

This is defined as

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where the O_i denote the observed cell counts and the E_i the expected cell counts.

The Deviance Goodness-of-Fit Test Statistics

Test a model against the saturated model using the deviance goodness-of-fit test, which is defined as follows:

$$G^2 = -2\sum_i O_i \log \frac{E_i}{O_i}$$

Under the null hypotheses, the deviance is also distributed as χ^2 with the appropriate degrees of freedom.

Significance of a test statistic is assessed by its *p*-value. Statistical significance is attained when the *p*-value is less than a predetermined minimum

level of significance, say α. The significance level α is often set at 0.05 or 0.01 (see Bishop, Fienberg, & Holland, 1989). Here the level α is set at 0.05.

In Table 2, the degrees of freedom of all the possible models for three-factor tables are listed. For more information about the model testing refer to Davis (1968), Kreiner (1987), and Landis, Heyman, and Koch (1978).

Analysis of Three-Factor Contingency Tables

Consider the different interaction models for three-factor tables and the mathematical formulation for the MLE of the expected counts (when it is possible) for each model.

Table 2. Degrees of freedom

Model	DF
[1][2][3]	$IJK - I - J - K + 2$
[12][3]	$(IJ - 1)(K - 1)$
[13][2]	$(IK - 1)(J - 1)$
[23][1]	$(JK - 1)(I - 1)$
[12][13]	$I(J - 1)(K - 1)$
[12][23]	$J(I - 1)(K - 1)$
[13][23]	$K(I - 1)(J - 1)$
[12][13][23]	$(I - 1)(J - 1)(K - 1)$
[123]	0

Complete Independence Model

This is the simplest model where all the factors are mutually independent and $u_{12} = u_{13} = u_{23} = u_{123} = 0$. The following different equivalent notations can be used to represent this model:

$$X \perp\!\!\!\perp Y \mid Z$$
$$\log\left(m_{ijk}\right) = u + u_1 + u_2 + u_3 \tag{7}$$
$$C = \{[1], [2], [3]\}$$

This model can be represented graphically as given in Figure 3.

Substitute the value of $\log(m_{ijk})$, as given in the equation (4) to the log-likelihood kernel as given by the Equation (6) and ignoring the constant term:

$$\log\left(\mathrm{f}\left(\left\{n_{ijk}\right\}\right)\right) = \sum_{ijk} n_{ijk} \log\left(m_{ijk}\right)$$
$$= \sum_{ijk} n_{ijk} \left(u + u_1 + u_2 + u_3\right)$$

After simplification, obtain

$$\mathrm{f}\left(\left\{n_{ijk}\right\}\right) = \exp\left(Nu + \sum_j u_1 n_{i..} + \sum_j u_2 n_{.j.} + \sum_k u_3 n_{..k} \right)$$

From the above expression, obtain the sufficient statistics for this models as marginal sub-tables: $C_1 = \{n_{i..}\}$, $C_2 = \{n_{.j.}\}$, and $C_3 = \{n_{..k}\}$, which are estimates of $m_{i..}$, $m_{.j.}$, and $m_{..k}$, respectively.

From equation (7), by summing over *jk*, *ik*, *ij*, and *ijk*, we obtain $m_{i..}$, $m_{.j.}$, $m_{..k}$, and $m_{...}$ as

$$\left\{m_{i..}\right\} = \exp\left(u + u_1\right) \sum_{jk} \exp\left(u_2 + u_3\right)$$
$$= \exp\left(u + u_1\right) \sum_j \exp\left(u_2\right) \sum_k \exp\left(u_3\right)$$
$$\left\{m_{.j.}\right\} = \exp\left(u + u_2\right) \sum_i \exp\left(u_1 + u_3\right)$$
$$= \exp\left(u + u_2\right) \sum_i \exp\left(u_1\right) \sum_k \exp\left(u_3\right)$$
$$\left\{m_{..k}\right\} = \exp\left(u + u_3\right) \sum_i \sum_j \exp\left(u_1 + u_2\right)$$
$$= \exp\left(u + u_3\right) \sum_i \exp\left(u_1\right) \sum_j \exp\left(u_2\right)$$
$$\left\{m_{...}\right\} = \exp\left(u\right) \sum_i \sum_j \sum_k \exp\left(u_1 + u_2 + u_3\right)$$
$$= \exp\left(u\right) \sum_i \exp\left(u_1\right) \sum_j \exp\left(u_2\right) \sum_k \exp\left(u_3\right)$$

From the above equations, get the expression for m_{ijk} as

$$m_{ijk} = \frac{m_{i..} m_{.j.} m_{..k}}{\left(m_{...}\right)^2}$$

Applying Birch's result, the estimates of m_{ijk} are

$$\hat{m}_{ijk} = \frac{n_{i..} n_{.j.} n_{..k}}{\left(n_{...}\right)^2}$$

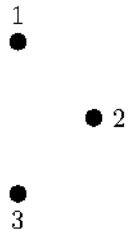


Figure 3. The complete independence model

Table 3. Personality type, cholesterol, and DBP marginal sub-tables of Table 1

Personality Type		Cholesterol		Diastolic Blood Pressure	
A	1027	Normal	1681	Normal	1928
B	1094	High	440	High	193

Table 4. Table of estimated cell counts for Example 4

		Diastolic Blood Pressure	
Personality Type	**Cholesterol**	**Normal**	**High**
A	Normal	739.90	74.07
	High	193.70	19.39
B	Normal	788.20	78.90
	High	206.30	20.65

Example 4: Consider the contingency table as given in Table 1. Under the complete independence assumption, the sufficient statistics are the marginal sub-tables given in Table 3. The table of fitted values, under the complete independence assumption, is given in Table 4. The G^2 statistic for the model is 8.723 (df: 4, p-value: 0.068), hence we conclude that the data supports the complete independence model. For details on the Chi-Squared test of independence, refer to Goodman (1971b).

Joint Independence Model

Under this model, two factors are jointly independent of the third factor. There are three versions of this model depending on which factor is unrelated to the other two. These three models are $(XY) \perp\!\!\!\perp Z$, $(XZ) \perp\!\!\!\perp Y$, and $(YZ) \perp\!\!\!\perp X$. Consider only $(XY) \perp\!\!\!\perp Z$ in detail as the others are comparable. Equivalent different notations are

$$\begin{aligned} &\log\left(m_{ijk}\right) = u + u_1 + u_2 + u_3 + u_{12} \\ &C = \left\{[12],[3]\right\} \end{aligned} \tag{8}$$

This model can also be represented graphically, as given in Figure 4.

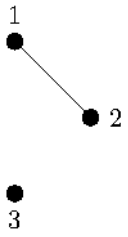


Figure 4. The joint independence model.

The sufficient statistics for this model are the marginal sub-tables $C_1 = \{n_{ij.}\}$ and $C_2 = \{n_{..k}\}$, which are the estimates of $m_{ij.}$ and $m_{..k}$. From equation (8), obtain

$$\begin{aligned} m_{ij.} &= \exp\left(u + u_1 + u_2 + u_{12}\right)\sum_k \exp\left(u_3\right) \\ m_{..k} &= \exp\left(u + u_3\right)\sum_i \sum_j \exp\left(u_1 + u_2 + u_{12}\right) \\ m_{...} &= \exp\left(u\right)\sum_i \sum_j \exp\left(u_1 + u_2 + u_{12}\right)\sum_k \exp\left(u_3\right) \end{aligned}$$

From the above equations, derive the closed form expression for m_{ijk} as

$$m_{ijk} = \frac{m_{ij.} m_{..k}}{m_{...}}$$

and, applying Birch's criteria,

$$\hat{m}_{ijk} = \frac{n_{ij.} n_{..k}}{n_{...}}$$

If the previous model of the complete independence $X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$ fits a data set, then the model, $(XY) \perp\!\!\!\perp Z$ will also fit. But the smallest model will be preferred.

Example 5: Consider the contingency table displayed in Table 5 to discuss this model. The sufficient statistics are given in Table 6. Under the assumptions of this model, the table of the expected cell counts is given in Table 7. The G^2 statistic for this model is 5.560 (df: 5, p-value: 0.351), hence we conclude that the data supports the joint independence model.

Table 5. Classroom behaviour table (Everitt, 1977)

		Risk	
Classroom Behaviour	**Adversity of School**	**Not at Risk**	**At Risk**
Nondeviant	Low	16	7
	Medium	15	34
	High	5	3
Deviant	Low	1	1
	Medium	3	8
	High	1	3

Table 6. Adversity*risk and classroom behaviour marginal sub-tables of Table 5

	Risk	
Adversity	**Not at Risk**	**At Risk**
Low	17	8
Medium	18	42
High	6	6

Classroom Behaviour	**Total**
Nondeviant	80
Deviant	17

Table 7. Table of estimated cell counts for Example 5

		Risk	
Classroom Behaviour	**Adversity of School**	**Not at Risk**	**At Risk**
Nondeviant	Low	14.020	6.597
	Medium	14.845	34.639
	High	4.948	4.948
Deviant	Low	2.979	1.402
	Medium	3.154	7.360
	High	1.051	1.051

Conditional Independence Model

Under this model, two factors are conditionally independent given the third factor. There are three version for this model as well, these are $X \perp\!\!\!\perp Y \mid Z$, $X \perp\!\!\!\perp Z \mid Y$, and $Y \perp\!\!\!\perp Z \mid X$. Consider only $X \perp\!\!\!\perp Y \mid Z$ in detail, as derivation for the others is similar. This model can be equivalently represented as

$$\begin{aligned} &\log\left(m_{ijk}\right) = u + u_1 + u_2 + u_3 + u_{13} + u_{23} \\ &C = \left\{[13],[23]\right\} \end{aligned} \tag{9}$$

The graph for this model is given in Figure 5.

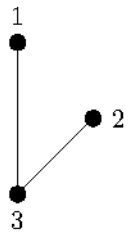


Figure 5. The conditional independence model

The sufficient statistics for this model are the marginal sub-tables $C_{13} = n_{i.k}$ and $C_{23} = n_{.jk}$, which are estimates of $m_{i.k}$ and $m_{.jk}$. From equation (9):

$$\begin{aligned} m_{i.k} &= \exp\left(u + u_1 + u_3 + u_{13}\right)\sum_j \exp\left(u_2 + u_{23}\right) \\ m_{.jk} &= \exp\left(u + u_2 + u_3 + u_{23}\right)\sum_i \exp\left(u_1 + u_{13}\right) \\ m_{..k} &= \exp\left(u + u_3\right)\sum_i \exp\left(u_1 + u_{13}\right)\sum_j \exp\left(u_2 + u_{23}\right) \end{aligned}$$

From the above three equations, obtain the closed form expression for m_{ijk} as

$$m_{ijk} = \frac{m_{ij.} m_{.jk}}{m_{..k}}$$

As before, applying Birch's criteria derive the expected counts for each cell as

$$\hat{m}_{ijk} = \frac{n_{ij.} n_{.jk}}{n_{..k}}$$

Example 6: Consider Table 8, infant's survival data taken from Bishop (1969). Assuming pre-natal care and survival are independent given a clinic, the sufficient statistics are given in Table 9. The G^2 statistic for this model is 0.082 (df: 2, p-value: 0.959), hence we conclude that the data supports the conditional independence model.

Table 8. Infant survival table

		Infant's Survival	
Clinic	**Pre-natal care**	**Died**	**Survived**
A	Less	3	176
	More	4	293
B	Less	17	197
	More	2	23

Table 9. Survival*clinic, clinic*pre-natal care, and clinic marginal sub-tables of Table 8

	Infant's Survival	
Clinic	**Died**	**Survived**
A	7	469
B	19	220

	Pre-natal Care	
Clinic	**Less**	**More**
A	179	297
B	214	25

Clinic	**Total**
A	476
B	239

Table 10. Table of estimated cell counts for Example 6

		Infant's Survival	
Clinic	**Pre-natal care**	**Died**	**Survived**
A	Less	2.632	176.367
	More	4.367	292.632
B	Less	17.012	196.987
	More	1.987	23.012

Uniform Association Model

This model is also known as the no three-factor interaction model, where $u_{123} = 0$. For this model the log-linear notation is [12] [13] [23], but there is no graphical representation for this model. Unlike the previous models, there are no closed-

form estimates for the expected cell counts/probabilities under this model. The MLEs can be computed by iterative procedures such as Iterative Proportional Fitting (IPF) and the Newton-Raphson method.

Example 7: Consider Table 11, auto accident data taken from Fienberg (1970). None of the models discussed in previous sections fit the data. Use the IPF algorithm to obtain the table of estimated counts as given in the Table 12. The G^2 statistic for this model is 0.043 (df: 1, p-value: 0.835), hence we conclude the data supports the marginal association model. For more information on IPF, refer to Deming and Stephan (1940) and Fienberg (1970). The IPF procedure implemented in the R package cat was used, available at cran.r-project.org.

Table 11. Auto accident data table

		Injury	
Accident Type	**Driver Ejected**	**Not Severe**	**Severe**
Collision	No	350	150
	Yes	26	23
RollOver	No	60	112
	Yes	19	80

Table 12. Table of estimated cell counts for Example 7

		Injury	
Accident Type	**Driver Ejected**	**Not Severe**	**Severe**
Collision	No	350.48858	149.51130
	Yes	25.51142	23.48870
RollOver	No	59.51104	112.48921
	Yes	19.48896	79.51079

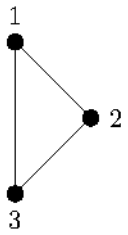


Figure 6. The saturated model

Saturated Model

For this model, the log-linear notation is [123]. In this case there is no independence relationship between the three factors. The expected cell counts are the same as the observed cell frequencies, e.g. $\hat{m}_{ijk} = n_{ijk}$. Graphical representation for the saturated model is given in Figure 6.

Example 8: Consider Table 13, a partial table which is based on clinical trial data from Koch, Amara, Atkinson, and Stanish (1983). None of the models fit the data; we leave this for the reader to verify.

Table 13. Results of a clinical trial for the effectiveness of an analgesic drug

		Response		
Status	**Treatment**	**Poor**	**Moderate**	**Excellent**
1	Active	3	20	5
	Placebo	11	4	8
2	Active	3	14	12
	Placebo	6	13	5

Model Selection for Decomposable Models

Model selection is now discussed for the decomposable models only, as a non-decomposable graphical model can be reduced to a decomposable one by adding a minimal number of edges to the graph. For details on minimum triangulation, refer to Rose, Tarjan, and Lueker (1970) and Heggernes (2006).

Though decomposable models are a restricted family of GLLMs, selecting an optimal model from the class of decomposable graphical models is known to be an intractable problem. Most of all existing model selection algorithms are based on forward selection, backward elimination, or a combination of the both. There is a vast literature available for model selection and inference on graphical models, e.g. see Wainwright and Jordan (2008), Dahinden, Kalisch, and Bühlmann (2010), Goodman (1971a), Ravikumar, Wainwright, and Lafferty (2010), and Allen and Liu (2012).

The Wermuth's procedure starts with the saturated model, a single clique that includes all the two-factor effects as given in Figure 7. The vertices *a*, *b*, *c*, *d*, *e*, and *f* correspond to the factors Attendance, Sex, School, Agree, Subject, and Plans, respectively.

Consider the backward model selection procedure for a real data set called women and mathematics (WAM), used in Fowlkes, Freeny, and Landwehr (1988). Wermuth's (1976) backward elimination algorithm is used. The data are shown in the Table 14.

Graphical models are completely specified by their two-factor interactions. By the hierarchical principle, if a two-factor term is set to zero, then any higher-order term that contain that particular two-factor term will also be set to zero.

Table 14. The women and mathematics data table

		School	Suburban School			
		Sex	Female		Male	
Plan		Preference	Attend	Not	Attend	Not
College	Maths-Sciences	Agree	37	27	51	48
		Disagree	16	11	10	19
	Liberal arts	Agree	16	15	7	6
		Disagree	12	24	13	7
Job	Maths-Sciences	Agree	10	8	12	15
		Disagree	9	4	8	9
	Liberal arts	Agree	7	10	7	3
		Disagree	8	4	6	4

		School	Urban School			
		Sex	Female		Male	
Plan		Preference	Attend	Not	Attend	Not
College	Maths-Sciences	Agree	51	55	109	86
		Disagree	24	28	21	25
	Liberal arts	Agree	32	34	30	31
		Disagree	55	39	26	19
Job	Maths-Sciences	Agree	2	1	9	5
		Disagree	8	9	4	5
	Liberal arts	Agree	5	2	1	3
		Disagree	10	9	3	6

In the next step, all the $\binom{6}{2}$ two-factor interactions are considered for elimination. Fix a backward elimination cut off level, $\alpha = 0.05$. Among the two-factor interactions, the terms having the largest p-value are considered for elimination, but only if the p-value exceeds α. From the Table 15, choose the edge (bf) for deletion, and the resulting graphical model is [$abcde$] [$acdef$].

In the next step, consider the cliques [$abcde$] and [$acdef$]. The edges ac, ad, ae, cd, ce, and de are common to both the cliques; they are not considered for

elimination because elimination of such edges may result in a non-decomposable model. The candidate edges for deletion are *ab*, *bc*, *bd*, *be*, *af*, *cf*, *df*, and *ef*. Let us examine the *p*-values for these edges as in the Table 16.

Delete the edge (*af*); the resulting graphical model is [*abcde*] [*cdef*]. Similarly, in the next step, the edge (*ad*) gets deleted and the resulting graphical model becomes [*abce*] [*bcde*] [*cdef*] as given in Figure 8.

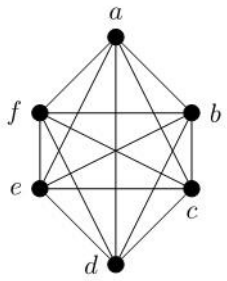


Figure 7. The saturated model for WAM

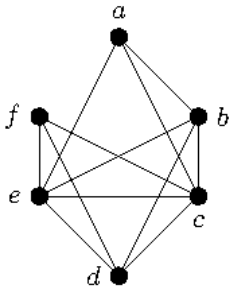


Figure 8. The fitted model for WAM

Table 15. WAM: [*abcde*]

Edge	Clique	d.f.	G^2	*p*-value
ab	[*acdef*] [*bcdef*]	16	18.585	0.29078
ac	[*acdef*] [*bcdef*]	16	20.689	0.19080
ad	[*acdef*] [*bcdef*]	16	14.172	0.58588
ae	[*acdef*] [*bcdef*]	16	18.781	0.28017
af	[*abcde*] [*bcdef*]	16	11.951	0.74734
bc	[*acdef*] [*abdef*]	16	26.739	0.04447
bd	[*acdef*] [*abcef*]	16	34.733	0.00432
be	[*acdef*] [*abcdf*]	16	56.570	0.00000
bf	[*acdef*] [*abcde*]	16	11.673	0.76616
cd	[*abcef*] [*abdef*]	16	29.439	0.02114
ce	[*abcdf*] [*abdef*]	16	26.052	0.05329
cf	[*abcde*] [*abdef*]	16	81.657	0.00000
de	[*abcdf*] [*abcef*]	16	78.248	0.00000
df	[*abcef*] [*abcde*]	16	46.221	0.00009
ef	[*abcde*] [*abcde*]	16	17.728	0.34005

Table 16. WAM: [*abcde*] [*acdef*]

Edge	Clique	d.f.	G^2	*p*-value
ab	[*bcde*] [*acdef*]	8	12.456	0.13198
bc	[*acde*] [*acdef*]	8	18.097	0.02051
bd	[*acde*] [*acdef*]	8	27.358	0.00061
be	[*acde*] [*acdef*]	8	49.723	0.00000
af	[*abcde*] [*cdef*]	8	5.822	0.66711
cf	[*abcde*] [*adef*]	8	73.014	0.00000
df	[*abcde*] [*acef*]	8	38.845	0.00001
ef	[*abcde*] [*acdf*]	8	10.881	0.20852

Table 17. WAM: [*abce*] [*bcde*] [*cdef*]

Edge	Clique	d.f.	G^2	*p*-value
ab	[*ace*] [*bce*] [*bcde*] [*cdef*]	4	10.606	0.03137
ac	[*bce*] [*ace*] [*bcde*] [*cdef*]	4	10.432	0.03374
ae	[*bce*] [*abc*] [*bcde*] [*cdef*]	4	10.426	0.03383
bd	[*abce*] [*cde*] [*bce*] [*cdef*]	4	25.507	0.00004
cf	[*abce*] [*bcde*] [*def*] [i]	4	67.832	0.00000

In the next step, candidate edges for deletion are [*ab*], [*ac*], [*ae*], [*bd*], and [*cf*]. None of the *p*-values are greater than $\alpha = 0.05$ as given in Table 17. So, stop with the model [*abce*] [*bcde*] [*cdef*].

Computational Details

All the experimental results were carried out using R 3.1.3. For fitting LLMs, there are several function in R, for example glm() and loglin() in the stats library and loglm() in the MASS library. For model selection, dmod() and backward() functions implemented in the package gRim were used. All the packages used are available at http://CRAN.R-project.org/.

Conclusion

The fundamental mathematical and statistical theory of GLLM and its applications were discussed, restricted to the complete table to make the discussion simple, because the tables having zero entries require special treatment. See Christensen (1997) for analysis of contingency tables with zero cell counts.

The limitations and open problems in the use of GLLM for recursive relationships can be further explored.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley-Interscience. doi: 10.1002/0471249688

Allen, G. I., & Liu, Z. (2012, October). *A log-linear graphical model for inferring genetic networks from high-throughput sequencing data*. Paper presented at the 2012 IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, PA. doi: 10.1109/bibm.2012.6392619

Andersen, A. H. (1974). Multidimensional contingency tables. *Scandinavian Journal of Statistics, 1*(3), 115-127. Available from http://www.jstor.org/stable/4615563

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association, 65*(331), 1248-1255. doi: 10.2307/2284291

Bartlett, M. S. (1935). Contingency table interactions. *Supplement to the Journal of the Royal Statistical Society, 2*(2), 248-252. doi: 10.2307/2983639

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological), 25*(1), 220-233. Available from http://www.jstor.org/stable/2984562

Bishop, Y. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics, 25*(2), 383-400. doi: 10.2307/2528796

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1989). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Christensen, R. (1997). *Log-linear models and logistic regression* (2nd ed.). New York, NY: Springer. doi: 10.1007/b97647

Dahinden, C., Kalisch, M., & Bühlmann, P. (2010). Decomposition and model selection for large contingency tables. *Biometrical Journal, 52*(2), 233-252. doi: 10.1002/bimj.200900083

Darroch, J. N. (1962). Interactions in multi-factor contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological), 24*(1), 251-263. Available from http://www.jstor.org/stable/2983765

Darroch, J. N., Lauritzen, S. L., & Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics, 8*(3), 522-539. doi: 10.1214/aos/1176345006

Davis, L. J. (1968). Exact tests for 2 × 2 contingency tables. *The American Statistician, 40*(2), 139-141. doi: 10.2307/2684874

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological), 41*(1), 1-31. Available from http://www.jstor.org/stable/2984718

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427-444. doi: 10.1214/aoms/1177731829

Edwards, D. (2000). *Introduction to graphical modeling* (2nd ed.). New York, NY: Springer-Verlag. doi: 10.1007/978-1-4612-0493-0

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics, 41*(3), 901-917. doi: 10.1214/aoms/1177696968

Fienberg, S. E., & Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference, 137*(11), 3430-3445. doi: 10.1016/j.jspi.2007.03.022

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 222*(594-604), 309-368. doi: 10.1098/rsta.1922.0009

Fowlkes, E. B., Freeny, A. E., & Landwehr, J. M. (1988). Evaluating logistic models for large contingency tables. Journal of the American Statistical Association, 83(403), 611-622. doi: 10.2307/2289283

Glonek, G. F., Darroch, J. N., & Speed, T. P. (1988). On the existence of maximum likelihood estimators for hierarchical loglinear models. *Scandinavian Journal of Statistics, 15*(3), 187-193. Available from http://www.jstor.org/stable/4616100

Goodman, L. A. (1969). How to ransack social mobility tables and other kinds of cross-classification tables. *American Journal of Sociology, 75*(1), 1-40. doi: 10.1086/224743

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interaction among multiple classifications. *Journal of the American Statistical Association, 65*(329), 226-256. doi: 10.2307/2283589

Goodman, L. A. (1971a). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics, 13*(1), 31-66. doi: 10.2307/1267074

Goodman, L. A. (1971b). The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association, 66*(334), 339-344. doi: 10.2307/2283933

Haberman, S. J. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *The Annals of Statistics, 1*(4), 617-632. doi: 10.1214/aos/1176342458

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago, IL: University of Chicago Press.

Heggernes, P. (2006). Minimal triangulations of graphs: A survey. *Discrete Mathematics, 306*(3), 297-317. doi: 10.1016/j.disc.2005.12.003

Knoke, D., & Burke, P. J. (1980). *Log-linear models*. Beverly Hills, CA: Sage. doi: 10.4135/9781412984843

Koch, G. G., Amara, I., Atkinson, S., & Stanish, W. (1983). *Overview of categorical analysis methods*. Paper presented at SAS Users Group International ‘83, New Orleans, LA.

Kreiner, S. (1987). Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scandinavian Journal of Statistics, 14*(2), 97-112. Available from http://www.jstor.org/stable/4616054

Kreiner, S. (1998). Interaction model. In *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistics Review, 46*(3), 237-254. doi: 10.2307/1402373

Lang, J. B. (1996a). Maximum likelihood methods for a generalized class of log-linear models. *The Annals of Statistics, 24*(2), 726-752. doi: 10.1214/aos/1032894462

Lang, J. B. (1996b). On the comparison of multinomial and Poisson log-linear models. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 253-266. Available from http://www.jstor.org/stable/2346177

Lang, J. B., McDonald, J. W., & Smith, P. W. (1999). Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *Journal of the American Statistical Association, 94*(448), 1161-1171. doi: 10.2307/2669932

Lauritzen, S. L. (1996). *Graphical models* (2nd ed.). New York, NY: Oxford University Press, Inc.

Lauritzen, S. L., Speed, T. P., & Vijayan, K. (1984). Decomposable graphs and hypergraphs. *Journal of the Australian Mathematical Society, 36*(1), 12-29. doi: 10.1017/s1446788700027300

Meeden, G., Geyer, C., Long, J., & Funo, E. (1998). The admissibility of the maximum likelihood estimator for decomposable log-linear interaction models for contingency tables. *Communications in Statistics – Theory and Methods, 27*(2), 473-493. doi: 10.1080/03610929808832107

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370-384. doi: 10.2307/2344614

Pearl, J., & Paz, A. (1987). Graphoids: A graph based logic for reasoning about relevance relations. *Advances in Artificial Intelligence, 2*, 357-363.

Pearson, K. (1904). *Mathematical contributions to the theory of evolution*. London, UK: Dulau and Co.

Ravikumar, P., Wainwright, M. J., & Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics, 38*(3), 1287-1319. doi: 10.1214/09-aos691

Rose, D., Tarjan, R. E., & Lueker, G. (1976). Algorithmic aspects of vertex elimination on graphs. *SIAM Journal on Computing, 5*(2), 146-160. doi: 10.1137/0205021

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning. 1*(1-2), 1-305. doi: 10.1561/2200000001

Wermuth, N. (1976). Model search among multiplicative models. *Biometrics, 32*(2), 253-263. doi: 10.2307/2529496

West, D. B. (2000). *Introduction to graph theory* (2nd ed.). Cambridge, MA: MIT Press.

Zelterman, D. (2006). *Models for discreet data* (2nd ed.). New York, NY: Oxford University Press, Inc.

Appendix A: Graphical Log-Linear Models for Four-Way Tables

Table 18. Graphical log-linear models for four-way tables

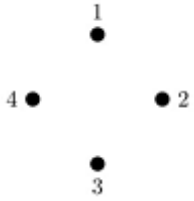
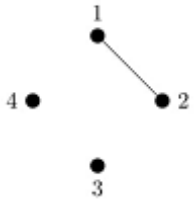
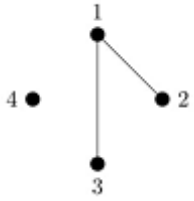
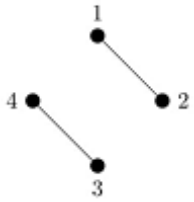
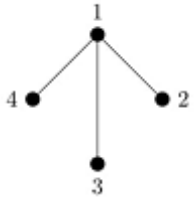
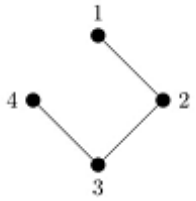
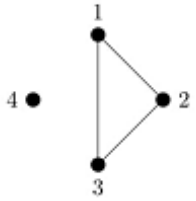
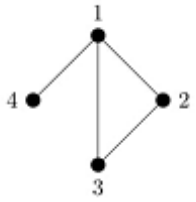
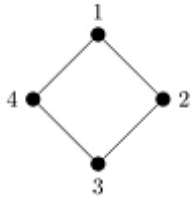
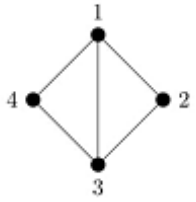
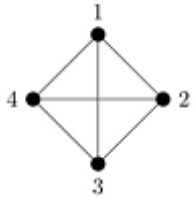
Model	Graph	Closed-Form Estimate
[1] [2] [3] [4]		$\hat{m}_{hijk} = \frac{n_{h\ldots} n_{.i..} n_{..j.} n_{\ldots k}}{n_{\ldots\ldots}^3}$
[12] [3] [4]		$\hat{m}_{hijk} = \frac{n_{hi..} n_{\ldots j.} n_{\ldots k}}{n_{\ldots\ldots}^2}$
[12] [13] [4]		$\hat{m}_{hijk} = \frac{n_{hi..} n_{h.j.} n_{\ldots k}}{n_{h\ldots} n_{\ldots\ldots}}$
[12] [34]		$\hat{m}_{hijk} = \frac{n_{hi..} n_{..jk}}{n_{\ldots\ldots}}$
[12] [13] [14]		$\hat{m}_{hijk} = \frac{n_{hi..} n_{h.j.} n_{h..k}}{n_{h\ldots}^2}$

Table 18, continued.

Model	Graph	Closed-Form Estimate
[12] [23] [34]		$\hat{m}_{hijk} = \frac{n_{hi..} n_{.ij.} n_{..jk}}{n_{.i..} n_{..j.}}$
[123] [4]		$\hat{m}_{hijk} = \frac{n_{hij.} n_{...k}}{n_{....}}$
[123] [14]		$\hat{m}_{hijk} = \frac{n_{hij.} n_{h..k}}{n_{h...}}$
[12] [23] [34] [14]		No closed-form estimate
[123] [134]		$\hat{m}_{hijk} = \frac{n_{hij.} n_{h.jk}}{n_{h.j.}}$
[1234]		$\hat{m}_{hijk} = n_{hijk}$

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 578-588.
doi: 10.22237/jmasm/1493598660

ISSN 1538 − 9472

Stochastic Model for Cancer Cell Growth through Single Forward Mutation

Jayabharathiraj Jayabalan
Pondicherry University
Puducherry, India

A stochastic model for cancer cell growth in any organ is presented, based on a single forward mutation. Cell growth is explained in a one-dimensional stochastic model, and statistical measures for the variable representing the number of malignant cells are derived. A numerical study is conducted to observe the behavior of the model.

Keywords: Mutation, probability generating function, differential-difference equation

Introduction

Cancer has complex and stochastic cell growth mechanisms. Malignant cancer cells arise from several mutations in the gene of a cell. It has been shown that a normal cell requires more than one stage to become a malignant cell (Tan & Brown, 1987). Stochastic growth is observed in malignant cells, and deterministic exponential growth is observed in normal cells. Most developed models are mixed, representing both deterministic and stochastic cell growth. The type of growth in a normal cell population depends on whether mutation has taken place.

Let $[X(t), t > 0]$ be a stochastic process denoting the number of normal cells in an organ at time t, and $[Y(t), t > 0]$ be a another stochastic process denoting the number of malignant cells in an organ at time t. Let us define a bivariate cell growth process $\{(X(t),Y(t)), t > 0\}$ representing the number of normal and malignant cells at time t. The growth of cells can be studied using the birth-and-death process. In literature, the process of malignant cell growth has been studied with homogeneous and non-homogeneous birth, death and mutation processes, but it seems most applicable when the study is conducted under a time-dependent

Jayabharathiraj Jayabalan is a Research Scholar in the Department of Statistics. Email at jayabharathi8@gmail.com.

environment. The birth-and-death process has been used to study the stochastic growth of a population, and the average and variance of the size of a population has been obtained for a given time period (Kendall, 1949). A similar approach can be applied to obtain the average and variance of the number of malignant cells in an organ at time t.

Assume there are x_0 number of normal cells and y_0 number of malignant cells in any organ at time $t = 0$ (initially). The model representing the cell division process for normal and malignant cells can be explained using either one or two variables. Consider a single forward mutation process for the transformation of normal cells into malignant cells, which reflects in the growth of the cell population. If a malignant cell is formed from a normal cell, and it remains in the same state till extinction, then there is not backward process of mutation. Let us assume that the normal cell has deterministic exponential growth; the expected number of cells in an organ at time t is then defined by (Serio, 1984),

$$\hat{X}(t) = x_0 \exp\left\{\int_0^t \left[\left(b_N(t) - d_N(t)\right) + m_{NM}(t)\right]dt\right\} \tag{1}$$

If it is assumed the malignant cells also show deterministic growth, i.e., the cell growth at the malignant stage is deterministic and exponential, then the expected number of malignant cells is as follows,

$$\hat{Y}(t) = y_0 \exp\left\{\int_0^t \left(b_M(t) - d_M(t)\right)dt\right\} \tag{2}$$

Assumptions

The model is developed based on the following assumptions:

1. Let the growth rate of normal cells from normal cells be $b_N(t)$, and the probability of growth of normal cells from normal cells in dt be $b_N(t)dt + o(dt)$. Let the death rate of normal cells be $d_N(t)$ and probability of the death of normal cells in dt be $d_N(t)dt + o(dt)$.

2. Let the growth rate of malignant cells from the malignant cells be $b_M(t)$, and probability of growth of malignant cells from malignant cells be $b_M(t)dt + o(dt)$. Let the death rate of malignant cells be $d_M(t)$

and probability of the death of malignant cells in dt be $d_M(t)dt + o(dt)$.

3. Let the growth rate of the normal cell population be represented by $\{(b_N(t) - d_N(t)) + \mu t_{NM}(t)\}$, and the growth rate of the malignant cell population by $(b_M(t) - d_M(t))$.

4. In a very small interval $(t + dt)$, let the probability of a mutation which transforms a normal cell into a malignant cell be $x\mu t_{NM}(t)dt + o(dt)$, where $X(t) = x_0$ at $t = 0$.

When a mutation takes place in a normal cell population, the number of normal cells is decreased by one and the number of malignant cells is increased by one. Assume that the growth rate for normal and malignant cell populations are different. For any organ, a certain number of cells is required for normal, proper functioning; normal functioning of any organ depends upon the number of cells.

The expected population size at time t can be described as $X(t) + Y(t)$. Assuming a deterministic growth for normal and malignant cells, then

$$\begin{aligned} T_c &= \hat{X}(t) + \hat{Y}(t) \\ &= x_0 \exp\left\{\int_0^t \left[\left(b_N(t) - d_N(t)\right) + m_{NM}(t)\right]dt\right\} \\ &\quad + y_0 \exp\left\{\int_0^t \left(b_M(t) - d_M(t)\right)dt\right\} \end{aligned} \tag{3}$$

where $X(t) = x_0$ and $Y(t) = y_0$ at $t = 0$. Hence, the number of normal cells in the population of an organ at time t is as follows

$$\begin{aligned} \hat{X}(t) &= x_0 \exp\left\{\int_0^t \left[\left(b_N(t) - d_N(t)\right) + m_{NM}(t)\right]dt\right\} \\ &\quad + y_0 \exp\left\{\int_0^t \left(b_M(t) - d_M(t)\right)dt\right\} - \hat{Y}(t) \end{aligned} \tag{4}$$

Assuming the above relation holds, there exists a stochastic dependence between $X(t)$ and $Y(t)$. There is no need to observe the variables $X(t)$ and $Y(t)$ as a

two-dimensional stochastic process; it is enough to consider one-dimensional stochastic process for the malignant cell population $[Y(t), t > 0]$ with the above relation. The above discussions deal with a non-homogeneous environment, and look more complex in mathematical derivations. For simplicity, let us assume a homogeneous environment with respect to birth, death and mutation parameters.

Stochastic Model

Let $f_M(y,t) = P\{Y(t) = y\}$ denote the probability density function of $Y(t)$. Assume that $f_M(y,t)$ exists and is differentiable with respect to both y and t; from Assumption 4, obtain the following relation (Armitage, 1952):

$$\begin{aligned} P\left[y < Y\left(t + dt\right) < y + dy \right] &= f\left(y, t + dt\right) dy \\ &= \left(1 - \mu_{NM} x\right) \left[1 - \left[\begin{array}{l} \left(b_N - d_N\right) \\ + \left(b_M - d_M\right) \\ + \mu_{NM} \end{array} \right] \right] dt \\ & f_M \left(y - \left[\begin{array}{l} \left(b_N - d_N\right) \\ + \left(b_M - d_M\right) \\ + \mu_{NM} \end{array} \right] y dt, t \right) dy \\ & + \mu_{NM} \left(x + 1\right) f_M \left(y - 1, t\right) dy dt + \left(\left(O dt \right) \right) dy + o\left(dy\right) \end{aligned} \tag{5}$$

By passing the limit on both sides in above equation, we obtain the differential-difference equation in the form as follows

$$\begin{aligned} & \frac{\partial f_M\left(y,t\right)}{\partial t} + \left[\left(b_N - d_N\right) + \left(b_M - d_M\right) + \mu_{NM} \right] y \frac{\partial f_M\left(y,t\right)}{\partial t} \\ & = \left(x + 1\right) \mu_{NM} f_M \left(y - 1, t\right) - \left[\begin{array}{l} \left(b_N - d_N\right) \\ + \left(b_M - d_M\right) \\ + \mu_{NM} \\ + \mu_{NM} x \end{array} \right] f_M \left(y, t\right) \end{aligned} \tag{6}$$

By using the relations given in equation (1), the above equation becomes,

$$\frac{\partial f_M(y,t)}{\partial t} + \begin{bmatrix} (b_N - d_N) \\ +(b_M - d_M) \\ +m_{NM} \end{bmatrix} y \frac{\partial f_M(y,t)}{\partial y}$$

$$= m_{NM} \begin{bmatrix} x_0 \exp\{(b_N - d_N)t\} \\ +y_0 \exp\left\{\begin{bmatrix}(b_M - d_M) \\ +m_{NM}\end{bmatrix} t\right\} \\ -y+1 \end{bmatrix} f_M(y-1,t) \qquad (7)$$

$$-\left[(b_N - d_N) + (b_M - d_M) + m_{NM} + m_{NM} x\right] f_M(y,t)$$

$$\frac{\partial f_M(y,t)}{\partial t} + \begin{bmatrix} (b_N - d_N) \\ +(b_M - d_M) \\ +m_{NM} \end{bmatrix} y \frac{\partial f_M(y,t)}{\partial y}$$

$$= m_{NM} \begin{bmatrix} x_0 \exp\{[(b_N - d_N) + m_{NM}]t\} \\ +y_0 \exp\{(b_M - d_M)t\} \\ -y+1 \end{bmatrix} f_M(y-1,t) \qquad (8)$$

$$- \begin{bmatrix} (b_N - d_N) + (b_M - d_M) \\ + m_{NM} \\ + m_{NM} \begin{pmatrix} x_0 \exp\{[(b_N - d_N) + m_{NM}]t\} \\ +y_0 \exp\{(b_M - d_M)t\} - y \end{pmatrix} \end{bmatrix} f_M(y,t)$$

The boundary condition for the above equations is

$$f_M(y,t) = 0, \text{ for } y < 0, t \geq 0$$
$$\lim_{y \to \infty} y^i f_M(y,t) = 0, \text{ for all } i \geq 0, t \geq 0.$$

The interest is to obtain the statistical moments such as mean and variance of malignant cells for a given time *t*. The probability generating function is

$$\phi(s,t) = \int_{-\infty}^{\infty} s^y f(y,t)\,dy : 0 \le s \le 1 \tag{9}$$

The partial derivative $\phi(s,t)$ with respect to s exists and from the boundary condition,

$$\int_{-\infty}^{\infty} y s^y \frac{\partial f(y,t)}{\partial s}\,dy = -\left[f(s,t) + s \log s \frac{\partial f(s,t)}{\partial s} \right]. \tag{10}$$

Multiply both sides of equation (1) by s^y and integrate, which yields the following differential equation for the generating function as

$$\begin{aligned}
&\int_{-\infty}^{\infty} s^y \frac{\partial f(y,t)}{\partial t}\,dy + \int_{-\infty}^{\infty} \left[\left(b_N - d_N\right) + \left(b_M - d_M\right) + \nu_{NM} \right] y s^y \frac{\partial f(y,t)}{\partial s}\,dy \\
&\quad = \int_{-\infty}^{\infty} \nu_{NM} \begin{bmatrix} x_0 \exp\left\{ \left[\left(b_N - d_N\right) + \nu_{NM} \right] t \right\} \\ + y_0 \exp\left\{ \left(b_M - d_M\right) t \right\} \\ - x + 1 \end{bmatrix} s^y f(y-1,t) \\
&\quad - \int_{-\infty}^{\infty} \left\{ \begin{bmatrix} \left[\left(b_N - d_N\right) + \left(b_M - d_M\right) + \nu_{NM} \right] \\ + \nu_{NM} \begin{bmatrix} x_0 \exp\left\{ \left[\left(b_N - d_N\right) + \nu_{NM} \right] t \right\} \\ + y_0 \exp\left\{ \left(b_M - d_M\right) t \right\} \end{bmatrix} - \nu_{NM} y \end{bmatrix} \right\} s^y f(y,t)
\end{aligned} \tag{11}$$

$$\frac{\partial f(s,t)}{\partial t} + \left[\left(b_N - d_N\right) + \left(b_M - d_M\right) + \mu_{NM}\right]\left\{-\left[f(s,t) + s\log s\frac{\partial f(s,t)}{\partial s}\right]\right\}$$
$$= \int_{-\infty}^{\infty} \mu_{NM}\begin{bmatrix} x_0\exp\left\{\left[\left(b_N - d_N\right) + \mu_{NM}\right]t\right\} \\ +y_0\exp\left\{\left(b_M - d_M\right)t\right\} \\ +1\end{bmatrix} s^y f(y-1,t)$$
$$- \int_{-\infty}^{\infty} \mu_{NM} y s^y f(y-1,t) \qquad (12)$$
$$- \int_{-\infty}^{\infty}\left\{\begin{bmatrix}\left(b_N - d_N\right) + \left(b_M - d_M\right) + \mu_{NM}\end{bmatrix} + \mu_{NM}\begin{bmatrix} x_0\exp\left\{\left[\left(b_N - d_N\right) + \mu_{NM}\right]t\right\} \\ +y_0\exp\left\{\left(b_M - d_M\right)t\right\}\end{bmatrix}\right\} s^y f(y,t)$$
$$+ \int_{-\infty}^{\infty} \mu_{NM} y s^y f(y,t)$$

$$\frac{\partial f(s,t)}{\partial t} - \left[-\mu_{NM}s^2 + \mu_{NM}s + \left[\left(b_N - d_N\right) + \left(b_M - d_M\right) + \mu_{NM}\right]s\log s\right]\frac{\partial f(s,t)}{\partial s}$$
$$= \mu_{NM}\left[x_0\exp\left\{\left[\left(b_N - d_N\right) + \mu_{NM}\right]t\right\} + y_0\exp\left\{\left(b_M - d_M\right)t\right\}\right](s-1)f(s,t) \qquad (13)$$

$$\frac{\partial f(s,t)}{\partial t} - \left[\mu_{NM}s + \left[\left(b_N - d_N\right) + \left(b_M - d_M\right) + \mu_{NM}\right]s\log s - \mu_{NM}s^2\right]\frac{\partial f(s,t)}{\partial s}$$
$$= \mu_{NM}\left[x_0\exp\left\{\left[\left(b_N - d_N\right) + \mu_{NM}\right]t\right\} + y_0\exp\left\{\left(b_M - d_M\right)t\right\}\right](s-1)f(s,t)$$

To obtain the moments, use the cumulant generating function of $y(t)$. Let $K(u,t) = \log\phi(s,t)$, where $s = e^u$. On simplification (Bharucha-Reid, 1960),

$$\frac{\partial K(s,t)}{\partial t} - \left[\mu_{NM} + \left[\left(b_N - d_N\right) + \left(b_M - d_M\right) + \mu_{NM}\right]u - \mu_{NM}e^u\right]\frac{\partial K(s,t)}{\partial s}$$
$$= \mu_{NM}\begin{bmatrix} x_0\exp\left\{\left[\left(b_N - d_N\right) + \mu_{NM}\right]t\right\} \\ +y_0\exp\left\{\left(b_M - d_M\right)t\right\}\end{bmatrix}\left(e^u - 1\right)K(s,t) \qquad (14)$$

Statistical Moments

The moments of the model can be obtained by expanding the cumulant generating function $K(u,t)$ on both sides of the expression as $K(u;t) = uE(Y(t)) + \frac{1}{2}u^2 Var(Y(t)) + L$, comparing the coefficient of the power of u's and v's, and equating coefficients on both sides of the equation. In this way we arrive at the following linear differential equations of constant parameters

$$\frac{d[E\{Y(t)\}]}{dt} = [(b_N - d_N) + (b_M - d_M)]E\{Y(t)\} + m_{NM}\begin{bmatrix} x_0 \exp\{[(b_N - d_N) + m_{NM}]t\} \\ + y_0 \exp\{(b_M - d_M)t\} \end{bmatrix} \quad (15)$$

$$\frac{d[Var\{Y(t)\}]}{dt} = 2[(b_N - d_N) + (b_M - d_M)]Var\{Y(t)\} - m_{NM}E\{Y(t)\} + m_{NM}\begin{bmatrix} x_0 \exp\{[(b_N - d_N) + m_{NM}]t\} \\ + y_0 \exp\{(b_M - d_M)t\} \end{bmatrix} \quad (16)$$

Solving the differential equation in (14) and (15) gives the average number of malignant cells and variance of number of malignant cells at a given time t. On solving above equation,

$$E[Y(t)] = C_1 \exp\{[(b_N - d_N) + (b_M - d_M)]t\} - \mu t_{NM}\left[\frac{x_0 \exp\{[(b_N - d_N) + \mu t_{NM}]t\}}{[(b_M - d_M) - \mu t_{NM}]} + \frac{y_0 \exp\{(b_M - d_M)t\}}{(b_N - d_N)}\right]$$

$$V\left[Y(t)\right] = C_2 \exp\left\{2\left[\left(b_N - d_N\right) + \left(b_M - d_M\right)\right]t\right\} - \frac{\mu_{NM} x_0 \left(b_M - d_M\right) \exp\left\{\left[\left(b_N - d_N\right) + \mu_{NM}\right]t\right\}}{\left[\left(b_M - d_M\right) + \left(b_N - d_N\right) - \mu_{NM}\right]\left[\left(b_M - d_M\right) - \mu_{NM}\right]} - \frac{\mu_{NM} y_0 \left[\left(b_N - d_N\right) + \mu_{NM}\right] \exp\left\{\left(b_M - d_M\right)t\right\}}{\left[\left(b_M - d_M\right) + 2\left(b_N - d_N\right)\left(b_N - d_N\right)\right]} + \frac{C_1 \exp\left\{2\left[\left(b_N - d_N\right) + \left(b_M - d_M\right)\right]t\right\}}{\left[\left(b_N - d_N\right) + \left(b_M - d_M\right)\right]}$$

The integration constants C_1 and C_2 will be obtained using the boundary conditions of the differential equations.

Numerical Study

For the fixed parameters and changing time, the changes are observed in the average, and expected and variance numbers of malignant cells in any organ are presented. The numerical study was conducted using Mathematica 8.0 software for solving the differential equations as given above in equations (14) & (15) numerically. The average and variance of number of malignant cells for fixed values of the parameters, $b_N = 0.0001$, $d_N = 0.0001$, $b_M = 0.04$, $d_M = 1.0 \times 10^{-7}$, $x_0 = 1.0 \times 10^5$, $y_0 = 1.0 \times 10^5$, and varying values of mutation rate and time are presented.

For the good maintenance of normal cell level, growth rate and death rate of normal cells are assumed to be equal, and large birth rate values for malignant cells and mutation rate are presented in the Figure 1. From the Figure, it is observed that there is a positive relationship between time and average number of malignant cells; a positive relationship between time and variance of number of malignant cells at lower values of mutation rates, and so on.

Conclusion

Birth, death, and single mutation processes with different growth rates are considered, to the study the growth of malignant cells by assuming $X(t)$ is dependent on $Y(t)$. The usual two dimensional models are replaced by a one dimensional model representing normal and malignant cells with interest in

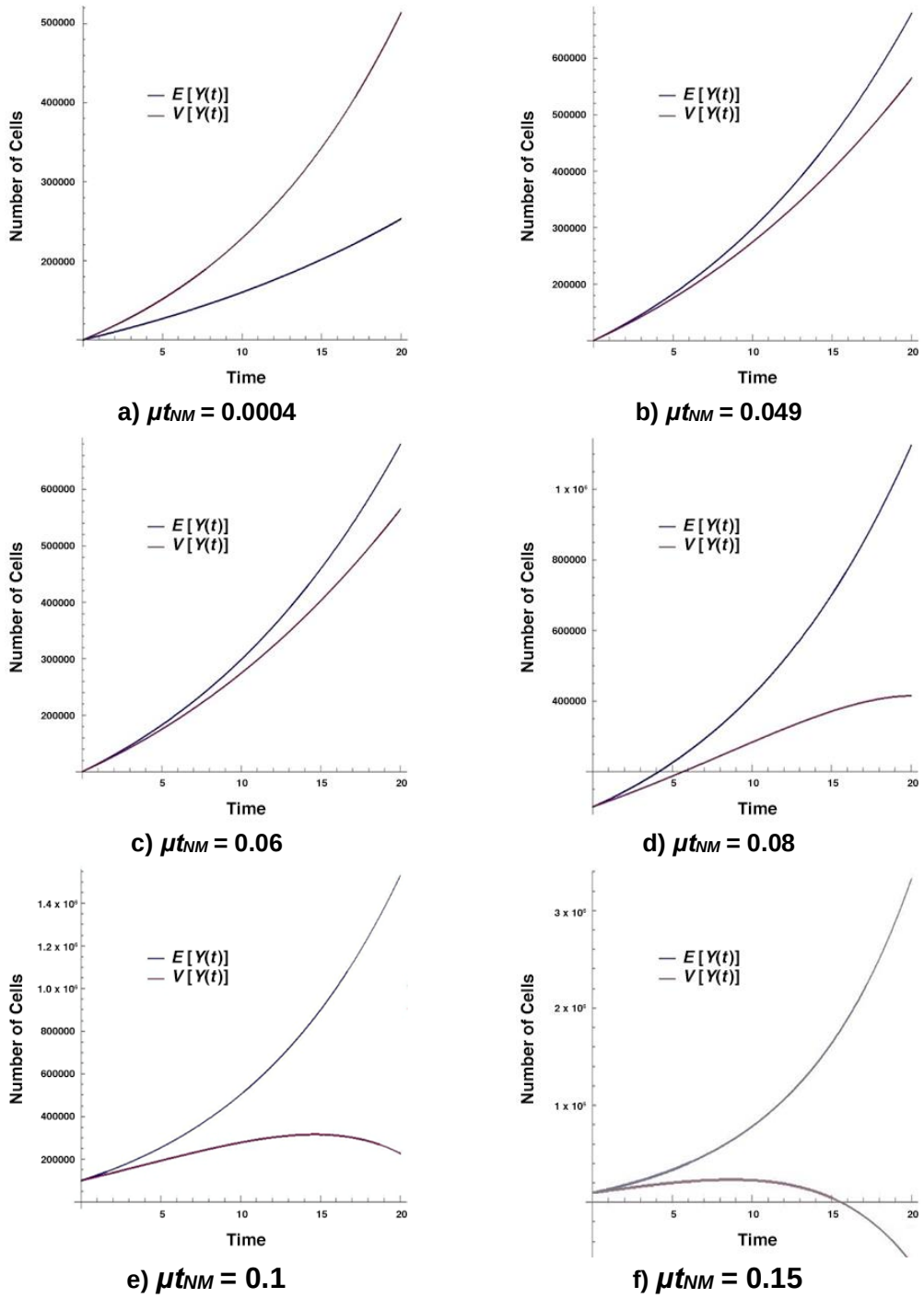


Figure 1. Variation in the moments with respect to time.

malignant cell population. The statistical measure shows that volatility of the malignant population decreases as the mutation rate increases, and average number of malignant cells increases drastically as the mutation rate increases. The results of this study may help to understand the behavior of malignant cells over a period of time with various decision parameters.

References

Armitage, B. P. (1952). The statistical theory of bacterial populations subject to mutation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1), 1–40.

Bharucha-Reid, A. T. (1960). *Elements of the theory of markov processes and their applications*. New York: McGraw-Hill.

Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological), 11*(2), 230-282.

Serio, G. (1984). Two-stage stochastic model for carcinogenesis with time-dependent parameters. *Statistics & Probability Letters, 2*(2), 95–103. doi: 10.1016/0167-7152(84)90057-9

Tan, W. Y. & Brown, C. C. (1987). A non-homogeneous two-stage carcinogenesis model. *Mathematical Modelling, 9*(8), 631-642. doi: 10.1016/0270-0255(87)90463-5

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 589-611.
doi: 10.22237/jmasm/1493598720

ISSN 1538 − 9472

An Empirical Comparison between Robust Estimation and Robust Optimization to Mean-Variance Portfolio

Epha Diana Supandi
State Islamic University, Sunan Kalijaga
Yogyakarta, Indonesia

Dedi Rosadi
Gadjah Mada University
Yogyakarta, Indonesia

Abdurakhman
Gadjah Mada University
Yogyakarta, Indonesia

Mean-variance portfolios constructed using the sample mean and covariance matrix of asset returns perform poorly out-of-sample due to estimation error. Recently, there are two approaches designed to reduce the effect of estimation error: robust statistics and robust optimization. Two different robust portfolios were examined by assessing the out-of-sample performance and the stability of optimal portfolio compositions. The performance of the proposed robust portfolios was compared to classical portfolios via expected return, risk, and Sharpe Ratio. The aim is to shed light on the debate concerning the importance of the estimation error and weights stability in the portfolio allocation problem, and the potential benefits coming from robust strategies in comparison to classical portfolios.

Keywords: Mean-variance portfolio, robust statistics, robust optimization

Introduction

The portfolio optimization approach proposed by Markowitz (1952) undoubtedly is one of the most important models in financial portfolio selection. This model is based upon the fundamental trade-off between expected return and risk, measured by the mean and standard deviation of return respectively. Therefore, Markowitz's model is called the mean-variance portfolio since this technique is highly reliant upon the value of a set of inputs, i.e. the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The goal of the portfolio allocation problem is to find weights $\mathbf{w}$ which represent the percentage of capital to be invested in each asset.

Epha Diana Supandi is a Lecturer in the Department of Mathematics. Email her at: epha.supandi@uin-suka.ac.id.

To compute the mean-variance portfolios, the mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ need to be estimated and both inputs are obtained from historical data. These estimators plug into an analytical or numerical solution to the investor's optimization problem. This leads to an important drawback in the mean-variance approach: the estimation error.

The fact that mean-variance "optimal" portfolios are sensitive to small changes in input data is well documented in the literature. Chopra and Ziemba (1993) showed that even slight changes to the estimates of expected return or risk can produce vastly different mean-variance optimized portfolios. Best and Grauer (1991) analyzed the sensitivity of optimal portfolios to changes in expected return estimates. Broadie (1993), meanwhile, showed how the estimated efficient frontier overestimates the expected returns of portfolios for various levels of estimation errors. Because of the ill effects of estimation errors on optimal portfolios, portfolio optimization has been called "error maximization" (see Michaud, 1989).

There are two standard methods extensively adopted in the literature to combat the impact of estimation error on portfolio selection. The first method is robust estimation, which can be quite robust to distributional assumptions. The introduction of robust estimation to portfolio optimization is relatively recent compared to the Markowitz foundational paper. Nevertheless, the subject has become very active in the last decade, as seen in the works of Lauprête (2001), Lauprete, Samarov, and Welsch (2002), Mendes and Leal (2003), Perret-Gentil and Victoria-Feser (2004), Welsch and Zhou (2007), and DeMiguel and Nogales (2009). The main difference among these approaches is in the term of the type of robust estimator used. Lauprête (2001) and Lauprete et al. (2002) used the least absolute deviation Huber estimator and trimean estimator, Mendes and Leal (2003) used the *M*-estimator, Perret-Gentil and Victoria-Feser (2004) used the *S*-estimator, Welsch and Zhou (2007) used the minimum covariance determinant estimator and Winsorization, and DeMiguel and Nogales (2009) used the *M*-estimator and the *S*-estimator. In their investigations, the portfolios constructed using a robust estimator outperformed those created using traditional mean-variance portfolio in the majority of cases.

The second method to deal with the estimation error is robust optimization. Robust portfolio optimization is a fundamentally different way of handling estimation error in the portfolio construction process. Unlike the previously-mentioned approaches, robust optimization considers the estimation error directly in the optimization problem itself. Introduced by Ben-Tal and Nemirovski (2002) for robust truss topology design, robust optimization is an emerging branch in the

field of optimization in which the solutions for optimization problems are obtained from uncertain parameters. The uncertainty is described using an uncertainty set which includes all, or most, possible realizations of the uncertain input parameters (see Pachamanova, Kolm, Fabozzi, & Focardi, 2007). The true mean and covariance matrix of asset returns lie in a fixed range. A robust portfolio, the one that optimizes the worst-case performance concerning with all possible values the mean vector and covariance matrix. The worst-case for robust optimization probably happened in the uncertainty sets (see, for example, Goldfarb & Iyengar, 2003; Tütüncü & Koenig, 2004; Engels, 2004; Garlappi, Uppal, & Wang, 2007; Lu, 2011).

The aim of this study is to shed light on the recent debate regarding the importance of the estimation error and weights' stability in the portfolio allocation problem and the potential benefits coming from robust portfolios in comparison to classical techniques. Here, two different robust portfolios have been investigated. The first portfolio was obtained by robust estimator to the mean-variance portfolio towards the *S*-estimators, constrained *M*-estimators, Minimum Covariance Determinant (MCD), and Minimum Volume Ellipsoid (MVE). The second one was obtained by robust optimization to the sample mean-variance portfolio where the formulation and the algorithm used in this paper were based on those developed by Tütüncü and Koenig (2004). We empirically compared two versions of robust asset allocation through the out-of-sample performance of those portfolio allocation approaches corresponding to the methodology of rolling horizon as proposed in DeMiguel and Nogales (2009).

The Mean-Variance Portfolio (Classical Portfolio)

It is assumed that the random vector $\mathbf{r} = (r_1, r_2, \ldots, r_N)'$ denotes random returns of the N risky assets with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A portfolio is defined to be a list of weights w_i for the assets $i = 1, \ldots, N$ that represent the amount of capital to be invested in each asset. We assumed that

$$\sum_{i=1}^{N} w_i = 1$$

meaning that capital is fully invested.

For a given portfolio $\mathbf{w}$, the expected return and variance were respectively given by: $E(\mathbf{w}'\mathbf{r}) = \mathbf{w}'\boldsymbol{\mu}$ and $Var(\mathbf{w}'\mathbf{r}) = \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$. Then, the classical mean-variance

portfolio models of Markowitz were formulated mathematically as the optimization problem:

$$\max_{\mathbf{w}} \mathbf{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}, \quad \text{s.t. } \mathbf{e}'\mathbf{w} = 1, \mathbf{w} \geq 0 \tag{1}$$

where $\boldsymbol{\mu} \in \Re^N$ is the vector of expected return, $\boldsymbol{\Sigma} \in \Re^{N \times N}$ is the covariance matrix of return, where $\Re^{N \times N}$ denotes the set of all $N \times N$ positive definite symmetric matrices, and $\mathbf{w} \in \Re^N$ is the vector of portfolio weight. The restriction $\mathbf{w} \geq 0$ means that short-selling is not allowed. The parameter γ can be interpreted as a risk aversion, since it takes into account the trade-off between risk and return of the portfolios.

The main criticisms against the Markowitz models centers on the observation that the optimal portfolios generated by this approach are often quite sensitive to the input parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. To make matters worse, these parameters can never be observed, and one has to settle for estimates found using some particular techniques.

Robust Portfolio Estimation

In this section, the class of portfolio policies based on the robust estimators is proposed where portfolio optimization and robust estimation are performed in two steps. It began by computing the robust estimators of the mean vector and covariance matrix of asset returns and followed by computing the portfolio policies by solving the classical minimum-variance problem (1), but replacing the sample mean and covariance matrix by their robust counterparts.

One of the most popular classes of robust estimators is affine equivariant robust estimators (see Maronna, Martin, & Yohai, 2007). Let $\left(\hat{\boldsymbol{\mu}}(\mathbf{r}), \hat{\boldsymbol{\Sigma}}(\mathbf{r})\right)$ be location and dispersion estimates corresponding to a sample = $(r_1, r_2, \ldots, r_N)'$. Then the estimates are affine equivariant if

$$\hat{\boldsymbol{\mu}}(\mathbf{A}\mathbf{r} + \mathbf{b}) = \mathbf{A}\hat{\boldsymbol{\mu}}(\mathbf{r}) + \mathbf{b} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}(\mathbf{A}\mathbf{r} + \mathbf{b}) = \mathbf{A}\hat{\boldsymbol{\Sigma}}(\mathbf{r})\mathbf{A}'$$

for any constant N-dimensional vector $\mathbf{b}$ and any non-singular $N \times N$ matrix $\mathbf{A}$. There are many different robust estimators for the mean and covariance in this class, such as S-estimators (Rousseeuw & Yohai, 1984), MVE and MCD proposed by Rousseeuw (1984), as well as CM-estimators (Kent & Tyler, 1996).

S-Estimators

S-estimators were first introduced (in the context of regression) by Rousseeuw and Yohai (1984). Later, they were applied to the multivariate scale and location estimation problem (Davies, 1992).

Let $\mathbf{r}$ be a data set in $\Re^N$. The *S*-estimators of the multivariate location $\hat{\boldsymbol{\mu}}(\mathbf{r}) \in \Re^N$ and scatter $\hat{\boldsymbol{\Sigma}}(\mathbf{r}) \in \Re^{N \times N}$ are defined as the solution to the problem of minimizing $|\boldsymbol{\Sigma}|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left[\left\{(\mathbf{r}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{r}_i - \boldsymbol{\mu})\right\}^{\frac{1}{2}}\right] = b_0 \tag{2}$$

where ρ denotes the loss function and b_0 satisfies $0 < b_0 < a_0 = \sup\{\rho\}$. As stated by Alqallaf (2003), it is natural to choose $b_0 = \mathrm{E}\left(\rho\left(\|\mathbf{r}\|\right)\right)$.

Let $\mathbf{r}$ be a data set in $\Re^N$ and $c_0 = b_0/\sup \rho$. If $c_0 \leq (n - N)/2n$, where $n \geq N + 1$, then the breakdown point $\varepsilon^* = \lceil nc_0 \rceil / n$, where $\lceil k \rceil$ denotes the nearest integer greater than or equal to k. The breakdown point for *S*-estimators is

$$\varepsilon^* = \frac{n - N + 1}{2n}$$

when

$$c_0 = \frac{(n - N)}{2n}$$

Portfolios based on *S*-estimators with biweight function were examined by Perret-Gentil and Victoria-Feser (2004) and, in a one-step approach, by DeMiguel and Nogales (2009).

CM-Estimators

As stated by Kent and Tyler (1996), the *CM*-estimator is defined via the minimization of an objective function subject to some constraints. For the data set $\mathbf{r}$ we defined the *CM*-estimators of the multivariate location $\hat{\boldsymbol{\mu}}(\mathbf{r}) \in \Re^N$ and scatter $\hat{\boldsymbol{\Sigma}}(\mathbf{r}) \in \Re^{N \times N}$ to be any pair which minimized the objective function

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{r}) = \frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{d}_i) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \tag{3}$$

subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{d}_i) \leq \varepsilon\rho(\infty) \tag{4}$$

where $\mathbf{d}_i = (\mathbf{r}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{r}_i - \boldsymbol{\mu})$, ρ denotes the loss function, and $\varepsilon \in (0, 1)$ refers to the breakdown point. Kent and Tyler (1996) showed that the breakdown point of the *CM*-estimate for data $\mathbf{r}$ in general is

$$\varepsilon^* = \min\left(\left\lceil \frac{n\varepsilon}{n} \right\rceil, \left\lceil \frac{n(1-\varepsilon) - N}{n} \right\rceil\right)$$

Minimum Volume Ellipsoid (MVE) Estimators

Rousseeuw (1984) introduced a highly robust estimator, the MVE estimator, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ was taken to be the center of the minimum volume ellipsoid covering at least half of the observations, and $\boldsymbol{\Sigma}$ was an N by N matrix representing the shape of the ellipsoid.

This approach attempted to seek the ellipsoid with the smallest volume covering h data points where $n/2 \leq h \leq n$. Formally, the estimate is defined as these $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ that minimized $|\boldsymbol{\Sigma}|$ subject to

$$\#\left\{i; (\mathbf{r}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{r}_i - \boldsymbol{\mu}) \leq c^2\right\} \geq \left\lceil \frac{n + N + 1}{2} \right\rceil \tag{5}$$

The constant c is chosen as $\chi^2_{N,0.5}$ and # denotes the cardinality. Portfolios based on MVE estimators were used by Kaszuba (2013). Let $\mathbf{r}$ be a data set in $\Re^N$ with $N \geq 2$, and let $n \geq N + 1$; then the breakdown point of MVE is

$$\varepsilon^* = \frac{\lfloor (n - N + 1)/2 \rfloor}{n}$$

Minimum Covariance Determinant (MCD) Estimators

The MCD estimators are highly robust estimators of multivariate location and scatter introduced by Rousseeuw (1984). Given an $n \times N$ data matrix $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)'$ with $\mathbf{r}_i = (\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iN})'$, it is focused on finding h (with $\lfloor (n + N + 1)/2 \rfloor \leq h \leq n$) observations whose classical covariance matrix has the lowest possible determinant. Then, the MCD estimator of location is the average of these h points, whereas the MCD estimator of scatter is their covariance matrix.

In 1999, Rousseeuw and Van Diressen constructed a very fast algorithm to calculate the MCD estimator. The new algorithm was called Fast-MCD based on the *C*-step. The Fast-MCD algorithm is defined as follows:

Algorithm 1. The Fast-MCD (Rousseeuw & Van Diressen, 1999)

1. Set an initial h-subset H_1, that is, beginning with a random $(N + 1)$-subset J.
2. Compute

$$\hat{\boldsymbol{\mu}}_0 = \frac{1}{N+1}\sum_{i \in \mathrm{J}} \mathbf{r}_i \quad \text{and} \quad \hat{\Sigma}_0 = \frac{1}{N+1}\sum_{i \in \mathrm{J}} \left(\mathbf{r}_i - \hat{\boldsymbol{\mu}}_0\right)\left(\mathbf{r}_i - \hat{\boldsymbol{\mu}}_0\right)'$$

   If $\left|\hat{\boldsymbol{\Sigma}}_0\right| = 0$, random observations are added to J until $\left|\hat{\boldsymbol{\Sigma}}_0\right| > 0$.
3. Apply the *C*-step to the initial h-subset H_1, and obtain the $\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1\right)$. If $\left|\hat{\boldsymbol{\Sigma}}_0\right| = 0$ or $\left|\hat{\boldsymbol{\Sigma}}_0\right| = \left|\hat{\boldsymbol{\Sigma}}_1\right|$, stop; otherwise, running another C-step produces $\left|\hat{\boldsymbol{\Sigma}}_2\right|$, and so on, until convergence is reached.

If the data are sampled from a continuous distribution, then these estimators have the breakdown point

$$\varepsilon^* = \min\left(\frac{n-h+1}{n}, \frac{h-p}{n}\right)$$

Portfolios based on MCD estimators were investigated by Zhou (2006), Welsch and Zhou (2007), and, in a modified version, by Mendes and Leal (2005).

S-estimators, *CM*-estimators, MVE, and MCD are used to construct robust portfolio mean-variance. A two-step approach to robust portfolio estimation is

proposed. First, compute a robust estimate of the mean vector and covariance matrix of asset returns. Second, solve the classical mean-variance problem (1), but replacing the sample mean and covariance matrix by their robust counterparts. Thus, given the robust estimators, the robust portfolio estimation can be found by solving the following optimization problem:

$$\max_{\mathbf{w}} \mathbf{w}'\hat{\boldsymbol{\mu}}_{\text{rob}} - \frac{\gamma}{2}\mathbf{w}'\hat{\boldsymbol{\Sigma}}_{\text{rob}}\mathbf{w}, \quad \text{s.t. } \mathbf{e}'\mathbf{w} = 1, \mathbf{w} \geq 0 \tag{6}$$

Robust Portfolio Optimization

Robust optimization has been developed to solve any problems related to the uncertainty in the decision environment and, therefore, sometimes it is referred to uncertain optimization (Ben-Tal & Nemirovski, 2002). Robust models have been adapted in portfolio optimization to resolve the sensitivity issue of the mean-variance portfolio to its inputs.

Robust portfolio optimization is to represent all available information about the unknown input parameters in the form of an uncertainty set that contains most of the possible values for these parameters.

Tütüncü and Koenig (2004) proposed a bootstrap method to determine the uncertainty sets. This method attempted to capture the uncertainty regarding the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in their uncertainty sets $\mathbb{U}_{\boldsymbol{\mu}}$ and $\mathbb{U}_{\boldsymbol{\Sigma}}$ by carrying out the following algorithm:

Algorithm 2. The construction of $\mathbb{U}_{\boldsymbol{\mu}}$ and $\mathbb{U}_{\boldsymbol{\Sigma}}$ using a block bootstrap method

1. Choose the block length (l). In our experiment, we used the non-overlapping block. Divide the data into n/l blocks in which block 1 became $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_l\}$ and block 2 became $\{\mathbf{r}_{l+1}, \mathbf{r}_{l+2}, \ldots, \mathbf{r}_{2l}\}$, ..., etc.
2. Resample the blocks and generate the bootstrap sample.
3. Compute the classical estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from bootstrap data.
4. Construct the empirical distribution of estimators by repeating step 2 and step 3 B times and sorting the bootstrap estimators from the smallest to largest ones.
5. Determine the $(1 - \alpha)100\%$ percent quintile of distribution of estimators

From algorithm 2, the uncertainty sets are defined as

$$\mathbb{U}_{\boldsymbol{\mu}} = \left\{ \boldsymbol{\mu} : \hat{\boldsymbol{\mu}}^{\mathrm{L}} \le \boldsymbol{\mu} \le \hat{\boldsymbol{\mu}}^{\mathrm{U}} \right\} \tag{7}$$

$$\mathbb{U}_{\boldsymbol{\Sigma}} = \left\{ \boldsymbol{\Sigma} : \hat{\boldsymbol{\Sigma}}^{\mathrm{L}} \le \boldsymbol{\Sigma} \le \hat{\boldsymbol{\Sigma}}^{\mathrm{U}}, \boldsymbol{\Sigma} \succeq 0 \right\} \tag{8}$$

Given the uncertainty sets of mean vector (7) and covariance matrix (8), then robust optimization (Rob.Opt) can be defined as follows:

$$\max_{\mathbf{w}} \mathbf{w}^{\mathrm{T}} \hat{\boldsymbol{\mu}}^{\mathrm{L}} - \frac{\gamma}{2} \mathbf{w}' \hat{\boldsymbol{\Sigma}}^{\mathrm{U}} \mathbf{w}, \quad \text{s.t.} \, \mathbf{e}^{\mathrm{T}} \mathbf{w} = 1, \mathbf{w} \ge 0 \tag{9}$$

Empirical Study

Data used in this study were collected from the Jakarta Stocks Exchange (JSE) consisting of 20 companies categorized as the blue chip. A blue chip is a stock in "a nationally recognized, well-established and financially sound company." ("Blue Chip", n.d.). Table 1 presents the list of companies.

The time series data span was from 04/02/2008 to 29/12/2014 with a total of 360 weekly returns. The first 260 observations (02/01/2008 to 07/01/2013) were used as the first window to perform the estimation and the uncertainty set. The last 100 observations (14/01/2013 to 29/12/2014) referred to the out-of-sample period and were used for the ex-post effectiveness analysis.

Table 1. Asset name for empirical analysis

No	Asset Name	No	Asset name
1	AALI = Astra Argo Lestari, Tbk	11	JSMR = Jasa Marga (Persero) Tbk
2	AKRA = Akr Corporindo Tbk	12	KLBF = Kalbe Farma Tbk
3	BBCA = Bank Centra Asia Tbk	13	LPKR = Lippo Karawaci Tbk
4	BBNI = Bank Negara Indonesia (Persero) Tbk	14	MNCN = Media Nusantara Citra Tbk
5	BBRI = Bank Rakyat Indonesia (Persero) Tbk	15	PGAS = Perusahaan Gas Neagara (Persero) Tbk
6	BMRI = Bank Mandiri (Persero) Tbk	16	PTBA = Tambang Batu Bara Asam (Persero) Tbk
7	CPIN = Charoen Pokphand Indonesia Tbk	17	SMGR = Semen Indonesia (Persero) Tbk
8	INDF = Indofood Sukses Makmur Tbk	18	TLKM = Telekomunikasi Indonesia (Persero) Tbk
9	INTP = Indocement Tunggal Prakarsa Tbk	19	UNTR = United Tractors Tbk
10	ITMG = Indo Tambangraya Megah Tbk	20	UNVR = Unilever Indonesia Tbk

Research Methodology

For an empirical analysis, several parameters have to be set. Firstly, for robust portfolio estimation, a translated biweight function is used as the loss function and the breakdown point is set at 45%. Meanwhile, in robust portfolio optimization, an important question is how to determine the uncertainty sets. The value α determines the most extreme parameter values that are still included in the uncertainty sets. The smaller α is, the larger an uncertainty set will be, and thus the greater the worst-case estimation errors will be. Hence, α can be interpreted as a parameter that captures the investor's tolerance for estimation errors (Fastrich & Winker, 2009). Therefore, to measure the level of sensitivity of the Rob.Opt model, set α = 0.05, 0.10, and 0.20.

Use the rolling-horizon procedure to compute the out-of-sample performance measures. This procedure has been implemented similarly as in DeMiguel and Nogales (2009). First, chose the window T = 260 to perform the estimation and the uncertainty sets. Second, using the return data in the estimation window, compute some optimal portfolio policies according to each strategy (classical portfolio, robust portfolio estimation, and robust portfolio optimization). Third, repeat the rolling-window procedure for the next month by including the four data points for the new date and dropping the four data points for the earliest period of the estimation window (we assumed that investors would rebalance their portfolios every one month). Continue this until the end of the dataset is reached. Therefore, at the end there is a time series of 25 portfolio weight vectors for each of the portfolios considered in the analysis.

The out-of-sample performance of each strategy was evaluated according to the following statistics: mean return, risk, Sharpe ratio, and portfolio turnover. Holding the portfolio $\mathbf{w}_t^s$ for one trading period gave the following out-of-sample excess return at time $t + 1$, that is $\hat{\mathbf{r}}_{t+1} = \mathbf{w}_t'^s \mathbf{r}_{t+1}^s$. After collecting the time series of 25 excess returns $\hat{\mathbf{r}}_{t+1}$, the out-of-sample mean return, standard deviation (risk), Sharpe ratio, and portfolio turnover are:

$$\hat{\mu}^s = \frac{1}{25}\sum_{t=1}^{25} \mathbf{w}_t'^s \mathbf{r}_{t+1}$$

$$\hat{\sigma}^s = \sqrt{\frac{1}{24}\sum_{t=1}^{25}\left(\mathbf{w}_t'^s \mathbf{r}_{t+1} - \hat{\boldsymbol{\mu}}^s\right)^2}$$

$$\mathrm{SR}^s = \frac{\hat{\mu}^s}{\hat{\sigma}^s}$$

$$\text{Turnover} = \frac{1}{24}\sum_{t=1}^{25}\sum_{j=1}^{5}\left(\left|w_{j,t+1} - w_{j,t}\right|\right)$$

where $w_{j,t}$ is the portfolio weight in asset j at time $t + 1$ but before rebalancing and $w_{j,t+1}$ is the desired portfolio weight in asset j at time $t + 1$. Therefore, the portfolio turnover is a measure of the variability in the portfolio holdings and can indirectly indicate the magnitude if the transaction costs associated to each strategy. Clearly, the smaller the turnover, the smaller the transaction costs associated to the implementation of the strategy.

Research Hypothesis

The research hypothesis is that the appropriate application of robust strategies in the construction of mean-variance portfolios allows the achievement of better investment results (measured with mean return and risk) in comparison to classical portfolios (benchmark). Hence, it is verified whether the given method allows one to obtain higher mean return compared to the classical method using the Wilcoxon signed rank test at significance level of 5%. Similarly, it is examined whether the robust methods will have lower risk (measured by standard deviation) compared to the classical method (see Kaszuba, 2013).

Results of Empirical Study

In the ninth column of Table 2, it can be observed that most of the return data were not normally distributed except AKRA, INTP, and UNVR. Also, UNVR had the best performance for having the highest mean return and the lowest risk (measured by standard deviation) compared to other stocks.

Presented in Table 3 are the out-of-sample performance of the classical and all robust approaches for each time window *win* in which the former serves as a benchmark. The results presented in Table 3 concern only portfolios for which risk aversion is equal to 10. Other risk aversion parameters were tested, such as $\gamma = 1$, 100, and 1000; the summary of these results are presented in Table 4.

It can be seen that the mean returns are higher in all seven robust approaches compared to the classical approach. An examination in the out-of-sample performance of portfolio returns indicated that the highest mean returns are obtained by robust portfolio estimation generated using *CM*-estimators (as presented in Table 3).

Table 2. Summary statistics of the 20 stocks used in the dataset

	Min	Max	Mean	Std. Dev	Var	Skew	Kurtosis	K.Smirnov
AALI	-0.4329	0.3459	-0.0007	0.0723	0.0052	-0.5740	6.4580	0.0004
AKRA	-0.2673	0.1982	0.0029	0.0612	0.0038	-0.1660	1.6770	0.4150
BBCA	-0.7071	0.1588	0.0017	0.0579	0.0034	-5.0540	62.0380	0.0004
BBNI	-0.4362	0.3920	0.0033	0.0634	0.0040	0.2190	11.6080	0.0002
BBRI	-0.6434	0.2975	0.0015	0.0655	0.0043	-2.6460	27.4440	0.0048
BMRI	-0.2744	0.2380	0.0033	0.0548	0.0030	-0.2460	4.2520	0.0293
CPIN	-1.5404	0.3868	0.0033	0.1109	0.0123	-7.4410	105.2200	0.0000
INDF	-0.2542	0.2654	0.0027	0.0556	0.0031	-0.1560	3.9300	0.0034
INTP	-0.4418	0.2747	0.0032	0.0579	0.0033	-0.9070	10.6050	0.0527
ITMG	-0.5557	0.3153	-0.0008	0.0773	0.0060	-0.9150	8.7970	0.0012
JSMR	-0.2942	0.1842	0.0036	0.0449	0.0020	-0.4700	6.3350	0.0409
KLBF	-1.5991	0.4970	0.0010	0.1038	0.0108	-10.0080	159.2970	0.0000
LPKR	-0.2587	0.3520	0.0011	0.0598	0.0036	0.4410	4.6370	0.0112
MNCN	-0.2801	0.5994	0.0032	0.0786	0.0062	1.3750	10.3110	0.0026
PGAS	-1.5549	0.2841	-0.0023	0.0974	0.0095	-11.3460	180.6230	0.0000
PTBA	-0.5771	0.2451	0.0002	0.0685	0.0047	-1.5180	14.2460	0.0007
SMGR	-0.6012	0.2766	0.0030	0.0591	0.0035	-2.5900	31.3040	0.0042
TLKM	-1.5864	0.1382	-0.0035	0.0930	0.0086	-13.7920	234.9950	0.0000
UNTR	-0.4215	0.2895	0.0009	0.0699	0.0049	-0.8050	7.9740	0.0018
UNVR	-0.1676	0.1436	**0.0042**	**0.0402**	0.0016	0.0500	1.5960	0.0590

Note: The bold values indicate the best performance of out-of-sample portfolio.

Also, it can be seen that MVE portfolios obtained higher Sharpe ratio than the ones obtained with the classical or other robust approaches. Whereas, in the context of risk, MCD generated using the fast algorithm exhibited the lowest risks. Meanwhile, MVE portfolios achieved the lowest turnover. Therefore, portfolio robust estimation (Rob.Est) created using a two-step approach (*CM*, *S*, MCD, and MVE portfolios) outperformed the classical approach for this case.

It can also be noticed that by analyzing the performance of Rob.Opt portfolios one can observe that increasing the investors' tolerance for estimation error α can decrease the performance of all out-of-sample for this portfolios.

Presented in Table 4 are the out-of-sample performance's portfolio, i.e., mean returns ($\hat{\mu}^s$), risk ($\hat{\sigma}^s$), Sharp Ratio (SR), and portfolio turnover (TO) at a number of different risk aversions, as well as different *p*-values of the Wilcoxon test for differences between the portfolios returns calculated with the given method and classical portfolios. The presented *p*-values for Wilcoxon test for observation pairs allows us to see whether the average weekly returns for the investigated portfolios were significantly higher than the average returns for classical portfolios.

Table 3. The out-of-sample performance of portfolio return for each time window *win* at $\gamma = 10$

		Rob.Est				Rob.Opt		
win	**Classic**	**CM**	**S**	**MCD**	**MVE**	***α*=5%**	***α*=10%**	***α*=20%**
1	0.0167	0.0641	0.0576	0.0523	0.0556	0.0080	0.0088	0.0077
2	0.0370	0.0450	0.0635	0.0477	0.0450	0.0485	0.0496	0.0500
3	-0.0017	0.0139	0.0135	0.0110	0.0099	-0.0032	-0.0033	-0.0037
4	0.0017	0.0060	0.0190	-0.0078	0.0151	0.0155	0.0137	0.0117
5	-0.0577	-0.0769	-0.0775	-0.0662	-0.0609	-0.0593	-0.0584	-0.0582
6	0.0434	0.0160	0.0317	0.0181	0.0087	0.0753	0.0742	0.0687
7	-0.0099	-0.0121	0.0012	-0.0030	0.0023	-0.0186	-0.0222	-0.0181
8	0.0684	0.1429	0.1275	-0.0030	0.1220	0.0248	0.0270	0.0314
9	0.0046	0.0310	0.0242	0.0178	0.0291	0.0139	0.0137	0.0122
10	0.0100	0.0313	0.0236	0.0266	0.0257	0.0043	0.0038	0.0045
11	-0.0122	-0.0247	-0.0228	-0.0238	-0.0237	-0.0155	-0.0144	-0.0149
12	0.0135	0.0277	0.0189	0.0229	0.0327	0.0088	0.0099	0.0118
13	-0.0281	0.0085	-0.0025	-0.0162	0.0050	-0.0166	-0.0167	-0.0227
14	0.0006	-0.0025	0.0074	-0.0051	-0.0017	-0.0090	-0.0083	-0.0070
15	0.0054	0.0056	-0.0193	0.0073	0.0083	0.0195	0.0186	0.0168
16	-0.0060	-0.0107	-0.0217	-0.0087	-0.0163	-0.0131	-0.0125	-0.0112
17	-0.0222	-0.0123	-0.0304	-0.0092	-0.0086	-0.0113	-0.0158	-0.0169
18	-0.0267	-0.0112	-0.0131	-0.0036	-0.0016	-0.0125	-0.0118	-0.0164
19	0.0006	0.0050	0.0079	0.0065	0.0060	0.0171	0.0149	0.0143
20	0.0389	0.0236	0.0223	0.0256	0.0309	0.0272	0.0285	0.0289
21	-0.0081	-0.0132	-0.0198	-0.0098	-0.0106	-0.0059	-0.0049	-0.0062
22	-0.0249	-0.0088	0.0136	-0.0112	0.0004	0.0103	0.0058	0.0024
23	-0.0248	-0.0398	-0.0594	-0.0338	-0.0479	-0.0171	-0.0181	-0.0193
24	-0.0011	0.0129	0.0079	0.0105	0.0093	0.0130	0.0109	0.0105
25	0.0200	0.0203	0.0456	0.0198	0.0037	0.0169	0.0150	0.0193
$\hat{\mu}^s$	0.0015	**0.0097**	0.0088	0.0026	0.0096	0.0048	0.0043	0.0038
$\hat{\sigma}^s$	0.0269	0.0396	0.0409	**0.0249**	0.0347	0.0257	0.0258	0.0256
SR	0.0555	0.2442	0.2142	0.1038	**0.2751**	0.1881	0.1678	0.1491
TO	1.5891	1.1840	**1.1097**	1.1527	1.2927	1.6178	2.0235	2.0165

Note: The bold values indicate the best performance

An examination in the out-of-sample performance of portfolio returns indicated that the highest mean returns were obtained by robust portfolios. Of the robust approaches, portfolios generated with *CM*-estimators achieved the higher mean returns at $\gamma = 1$ and 10. Meanwhile, Rob.Opt portfolios obtained higher mean returns at $\gamma = 100$ and 1000.

Table 4. Out-of-sample performance's portfolio i.e. mean returns ($\hat{\mu}^s$), risk ($\hat{\sigma}^s$), Sharpe ratio (SR) and portfolio turnover (TO) at different of risk aversions

			Rob.Est				Rob.Opt		
		Classic	**CM**	**S**	**MCD**	**MVE**	***α*=5%**	***α*=10%**	***α*=20%**
$\gamma = 1$	$\hat{\mu}^s$	-0.0076	**0.0160**	0.0128	0.0142	0.0159	-0.0003	-0.0033	-0.0073
	p-value	1.0000	0.0773	0.0773	0.1759	0.0954	0.4410	0.6169	0.8289
	$\hat{\sigma}^s$	0.0435	0.0572	0.0592	0.0752	0.0595	**0.0341**	0.0351	0.0385
	p-value	1.0000	0.0075*	0.001*	0.0012*	0.0274*	0.0004*	0.0000*	0.0000*
	SR	-0.1751	**0.2801**	0.2163	0.1891	0.2671	-0.0093	-0.0929	-0.1898
	TO	1.9026	1.9372	2.0000	1.8958	1.9282	**1.6731**	1.6461	2.0955
$\gamma = 10$	$\hat{\mu}^s$	0.0015	**0.0097**	0.0088	0.0026	0.0096	0.0048	0.0043	0.0038
	p-value	1.0000	0.3859	0.3350	0.5004	0.2887	0.5379	0.5900	0.7148
	$\hat{\sigma}^s$	0.0269	0.0396	0.0409	**0.0249**	0.0347	0.0257	0.0258	0.0256
	p-value	1.0000	0.0000*	0.0004*	**0.0000***	0.0000*	0.0000*	0.0000	0.0000*
	SR	0.0555	**0.2442**	0.2142	0.1038	0.2751	0.1881	0.1678	0.1491
	TO	1.5891	1.1840	**1.1097**	1.1527	1.2927	1.6178	2.0235	2.0165
$\gamma = 100$	$\hat{\mu}^s$	0.0058	0.0062	0.0054	0.0049	0.0064	**0.0067**	0.0066	0.0065
	p-value	0.9693	0.9540	0.8929	0.9234	0.9234	0.9234	0.9234	0.9234
	$\hat{\sigma}^s$	0.0290	0.0276	0.0267	**0.0252**	0.0262	0.0292	0.0290	0.0288
	p-value	1.0000	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*
	SR	0.2010	0.2239	0.2033	0.1953	**0.2449**	0.2308	0.2283	0.2273
	TO	1.4530	1.0739	**0.9222**	1.0650	1.0682	1.6414	2.0457	2.0276
$\gamma = 1000$	$\hat{\mu}^s$	0.0057	0.0042	0.0041	0.0045	0.0053	**0.0061**	0.0060	0.0060
	p-value	1.0000	0.8626	0.8929	0.8929	0.9693	0.9847	0.9847	0.9847
	$\hat{\sigma}^s$	0.0286	0.0243	0.0244	0.0245	**0.0240**	0.0290	0.0288	0.0290
	p-value	1.0000	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*
	SR	0.1978	0.1718	0.1678	0.1835	**0.2208**	0.2103	0.2083	0.2069
	TO	1.4585	**1.0547**	1.0112	1.0072	1.3808	2.0360	2.0317	2.0270

Note: The bold values indicate the best performance; an asterisk (*) indicates *p*-values at a significance level of 0.05

The corresponding results for the portfolio risk showed that the Rob.Est portfolios were better than two portfolio approaches (i.e. classical and Rob.Opt). The lowest portfolio risk was achieved by Rob.Est in the majority of the scenarios (γ = 10, 100 and 1000). The research demonstrated that portfolios generated with MCD and MVE achieved a lower portfolio risk compared to *S*- and *CM*-

estimators. Therefore, it is obvious if the largest Sharpe ratios are obtained by Rob.Est in all cases.

Comparing portfolio turnover values, one can observe that for all portfolios, increasing the risk aversion value from 1 to 1000 has caused these values to decrease. Portfolios created using robust estimators (*CM* and *S*) had the lowest turnover except at $\gamma = 1$.

An empirical study using the real market data indicated that, for all robust portfolios with robust estimation and robust optimization on portfolio weights, there were statistically significant improvements in the risk. The classical portfolios were characterized by a much higher risk than robust portfolios. However, in the context of mean return, the difference in performances between robust techniques and classical techniques did not seem to be statistically significant (p-value > 0.05), the robust estimation techniques were able to deliver more stability in the portfolio weights in comparison to the classical approach. The main implication of this finding is that, if we assume equal performance across techniques, investors will be better off by choosing a strategy that does not require any radical changes in the portfolio composition over time. These substantial changes in portfolio composition are rather difficult to be implemented in practice due to (i) management costs; and (ii) negative cognitive aspects perceived by investors and/or investment managers (see Santos, 2010).

Because the aim was to examine portfolios regarding their robustness properties, a small turnover indicates the stability of portfolio, which means it is more robust. From the point of view of an investor, the stability of weights in a portfolio constructed by them throughout the entire duration of the investment is a significant element. In this case, as seen in Table 4, the smallest turnover is achieved by Rob.Est. These findings are corroborated by the visual inspection of Figure 1 and Figure 2, which show the time-varying portfolio weights and boxplots of each portfolio technique.

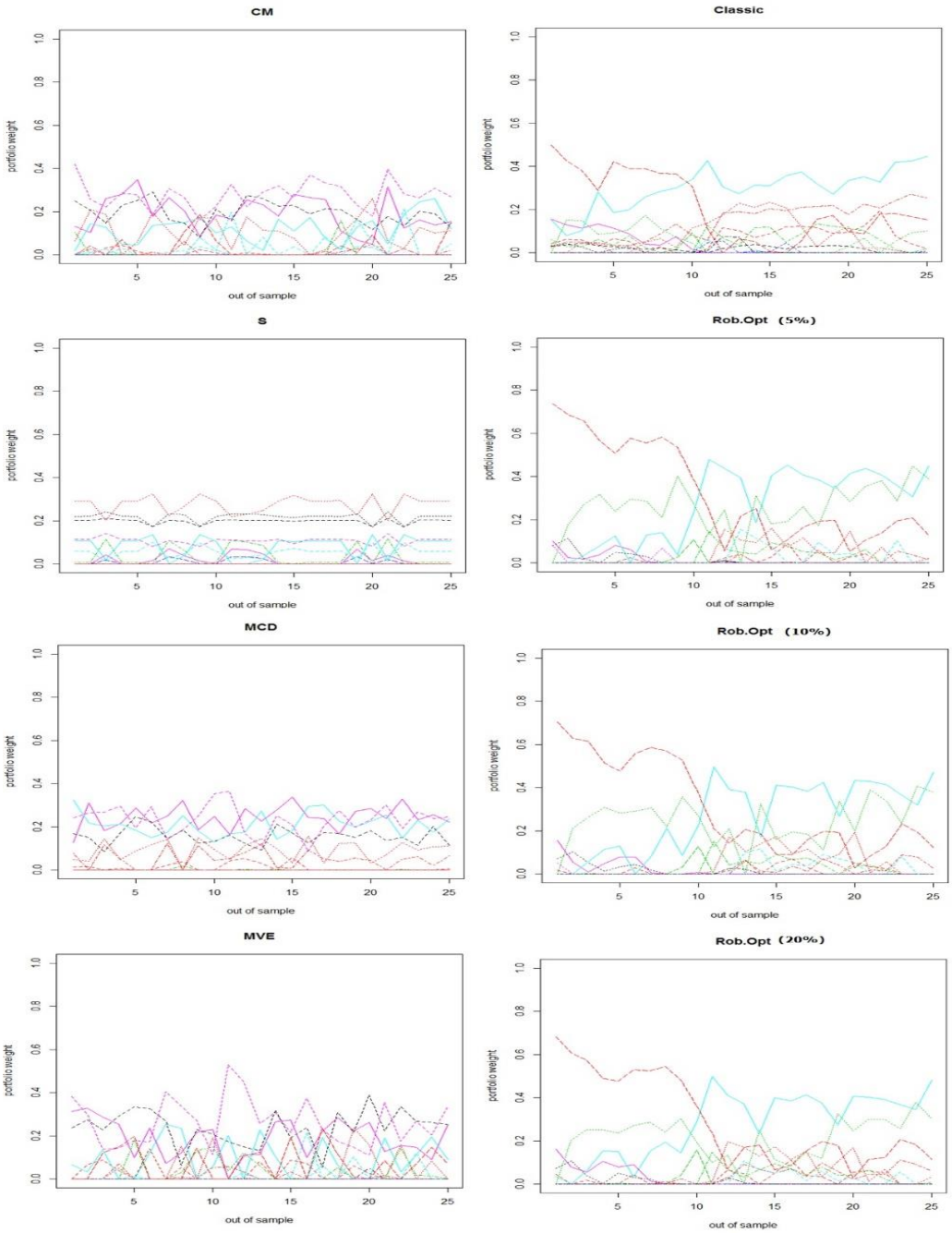


Figure 1. Time-varying portfolio weights for classical portfolio and robust portfolios for the case of $\gamma = 10$

Plotted in Figure 1 are the time-varying portfolio weights for the classical portfolio, robust portfolio optimization (right column of graphs), and robust portfolio estimation (left column of graphs) at risk aversion is equal to 10. All of the eight graphs map the time window *win* on the *x*-coordinate, while the *y*-coordinate maps the portfolio weights. Other risk aversion parameters were tested, such as $\gamma = 1$, 100, and 1000, but the insights from the results were similar, and thus the results are presented only for the case $\gamma = 10$.

It can be seen that Figure 1 corroborates the main findings by showing the high instability associated to the time-varying portfolio weights (compositions) of classic and Rob.Opt in contrast to the relative stability in the composition of Rob.Est.

Figure 2 gives the boxplots of the portfolio weights of classical portfolio, robust portfolio estimation, and robust portfolio optimization for the case of $\gamma = 10$.

Each graph in Figure 2 contains 20 boxplots corresponding to each of the twenty assets (for detail, see Table 1). Finally, the box for each portfolio weight has lines at the 25^{th}, 50^{th}, and 75^{th} percentile values of the portfolio weights. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data. Extreme portfolio weights that have values beyond the whiskers are also depicted (as indicated by the white circles). We have tested other risk aversion parameters, such as $\gamma = 1$, 100 and 1000, but the insights from the results were similar and thus the results are presented only for the case $\gamma = 10$.

It can be observed from Figure 2 that the mean-variance portfolios (classical and Rob.Opt) are much more unstable than the Rob.Est portfolios. For instance, for $\gamma = 10$, it can be seen that the Rob.Opt portfolios generated using $\alpha = 5\%$ concentrate the allocation in only five assets of twenty available, and the allocation between these five assets radically changed in the period analyzed (see the second row of the second column in Figure 2). This is reflected in the high portfolio turnover as achieved by Rob.Opt (2.0235). As in the previous strategy, the changes in the portfolio weights associated to the Rob.Est were more stable over time since it produced little turnover.

A further step in the analysis was to check which observations are considered outliers and were responsible for this instability of the portfolios. To do so, we used a diagnostic tool called Mahalanobis distance. Briefly, the Mahalanobis distance can identify which observations are quite far from the bulk of data to be considered outliers (Werner, 2003).

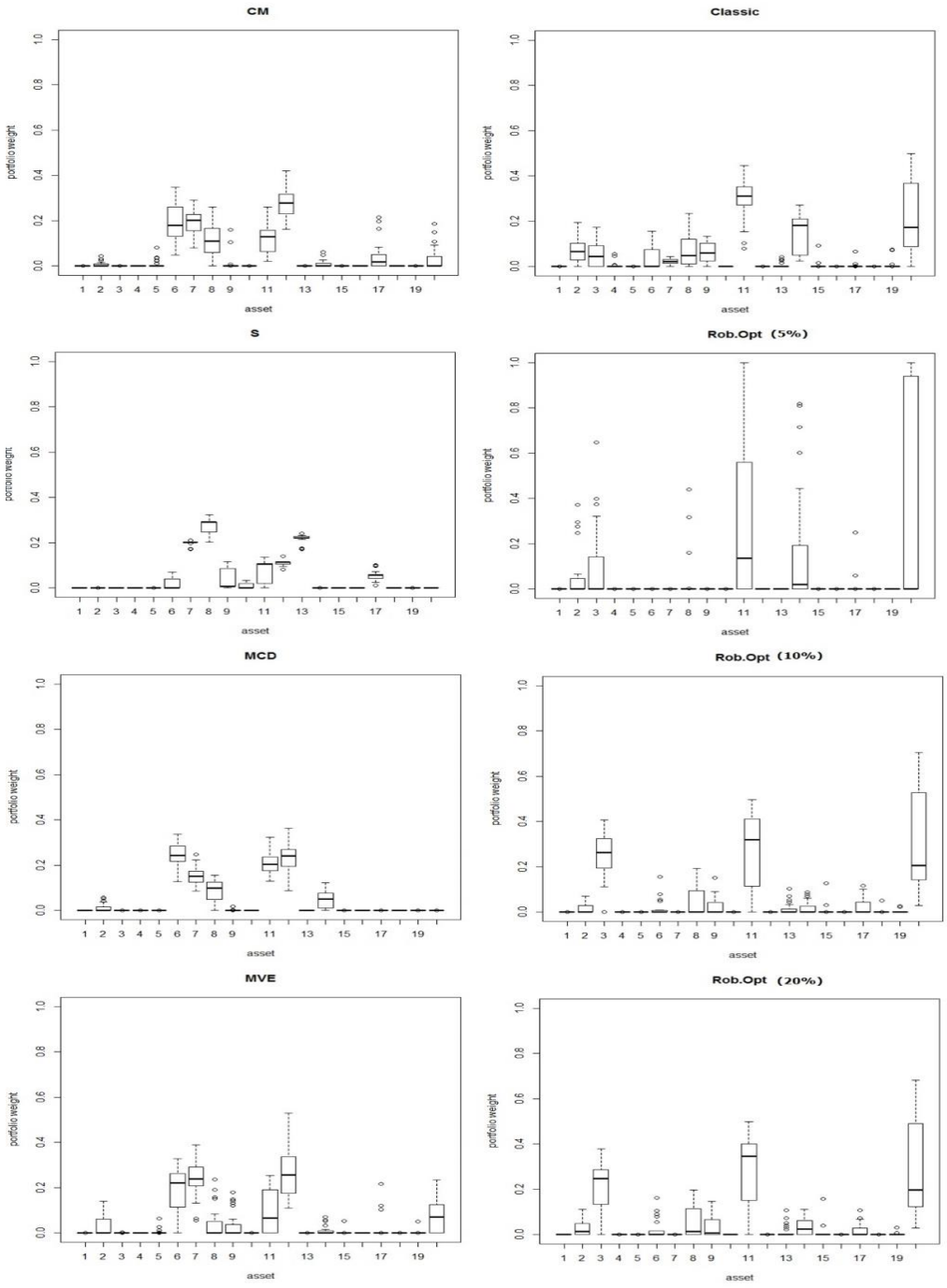


Figure 2. Boxplots of the portofolio weights for classical portfolio and robust portfolios for the case of $\gamma = 10$

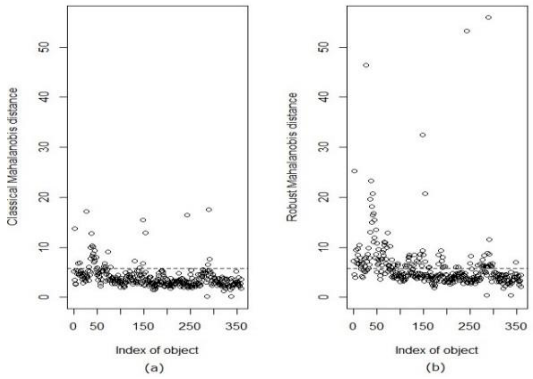


Figure 3. Mahalanobis distance of each of the 360 returns

Figure 3 shows the Mahalanobis distances of data using classical estimators in panel (a) and MCD estimators in panel (b). It is found that both pictures exhibited extreme return observations compared to the majority. They were detected to have a very strong influence on the classical estimates of the optimal portfolio weights (compositions). In short, it has been found that the outlying observations in the data have a strong influence on the composition of the resulting optimal portfolios.

In summary, the robust techniques lead to an improvement compared to the classical approach. Of the robust approaches, the robust estimation clearly outperforms the robust optimization approach. This improvement is possible due to the properties of robust estimator, which is not influenced by the presence of outliers.

Conclusion

In this work, two different robust techniques, robust estimation and robust optimization, have been empirically tested and compared with a classical approach. From the results presented in the previous section, some important implications for investment decisions based on portfolio selection policies can be pointed out.

Based on an empirical analysis, it is shown that the robust portfolio estimation (Rob.Est) significantly outperformed the classical portfolio and robust portfolio optimization in terms of out-of-sample performance, i.e. mean excess return, risk, Sharpe ratio, and portfolio turnover, in the majority of the scenarios. The portfolio compositions of Rob.Est are shown to be more stable and

consequently lead to a reduction of the transaction cost. This is simply because robustly estimated parameters will be closer to the true parameter values when there are some extreme observations (outliers) than their classical counterparts. Meanwhile, the portfolio compositions of Rob.Opt are heavily biased as this method works on a worst-case approach, so it can be detrimentally influenced by outliers in the data

Therefore, in this case, of the robust approaches the robust estimation clearly outperforms the robust optimization approach. In future research, the robust estimation should be combined with robust optimization in the formation of the optimal portfolio.

References

Alqallaf, F. A. (2003). *A new contamination model for robust estimation with large high-dimensional data sets* (Doctoral dissertation). University of British Columbia, British Columbia, Canada. Retrieved from https://www.stat.ubc.ca/~ruben/website/Fatemah_thesis.pdf

Ben-Tal, A., & Nemirovski, A. (2002). Robust optimization: Methodology and applications. *Mathematical Programming, 92*(3). 453-480. doi: 10.1007/s101070100286

Best, M. J., & Grauer, R. R. (1991). On the sensitivity of mean-variance efficient portfolios to changes in asset means: some analytical and computational results. *Review of Financial Studies, 4*(2), 315-342. doi: 10.1093/rfs/4.2.315

Blue chip. (n.d.) Retrieved from http://www.investopedia.com/terms/b/bluechip.asp

Broadie, M. (1993). Computing efficient frontiers using estimated parameters. *Annals of Operations Research, 45*(1), 21-58. doi: 10.1007/bf02282040

Chopra, V. K., & Ziemba, W. T. (1993). The effects of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management, 19*(2), 6-11. doi: 10.3905/jpm.1993.409440

Davies, P. L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics, 20*(4), 1828-1843. doi: 10.1214/aos/1176348891

DeMiguel, V., & Nogales, F. J. (2009). Portfolio selection with robust estimation. *Journal of Operation Research, 57*(3). 560-577. doi: 10.1287/opre.1080.0566

Engels, M. (2004). *Portfolio optimization: Beyond Markowitz* (Master's thesis). Universiteit Leiden, Leiden, Netherlands. Retrieved from http://web.math.leidenuniv.nl/scripties/Engels.pdf

Fastrich, B., & Winker, P. (2009). Robust Portfolio Optimization with a hybrid heuristic algorithm. *Computational Management Science, 9*(1), 63-88. doi: 10.1007/s10287-010-0127-2

Garlappi, L., Uppal, R., & Wang, T. (2007). Portfolio selection with parameter and model uncertainty: multi-prior approach. *Review of Financial Studies, 20*(1), 41-81. doi: 10.1093/rfs/hhl003

Goldfarb, D., & Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research, 28*(1), 1-38. doi: 10.1287/moor.28.1.1.14260

Kaszuba, B. (2013). Empirical comparison of robust portfolios' investment effects. *The Review of Finance and Banking, 5*(1), 47-61. Retrieved from http://www.rfb.ase.ro/articole/ARTICLE_IV.pdf

Kent, J. T., & Tyler, D. E. (1996). Constrained *M*-estimation for multivariate location and scatter. *The Annals of Statistics, 24*(3), 1346-1370. doi: 10.1214/aos/1032526973

Lauprête, G. J. (2001). *Portfolio risk minimization under departures from normality* (Doctoral disertation). Massachusetts Institute of Technology, Cambridge, MA. Retrieved from http://hdl.handle.net/1721.1/8303

Lauprete, G. J., Samarov, A. M., & Welsch, R. E. (2002). Robust portfolio optimization. *Metrika, 55*(1), 139-149. doi: 10.1007/s001840200193

Lu, Z. (2011). A computational study on robust portfolio selection based on a joint ellipsoidal uncertainty set. *Mathematical Programming, 126*(1), 193-201. doi: 10.1007/s10107-009-0271-z

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77-91. doi: 10.2307/2975974

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Chichester, England: John Wiley and Sons. doi: 10.1002/0470010940

Mendes, B. V. M., & Leal, R. P. C. (2003). *Robust multivariate modelling in finance* (COPPEAD working paper series, no. 355). Rio de Janeiro, Brazil: Federal University at Rio de Janeiro.

Mendes, B. V. M., & Leal, R. P. C. (2005). Robust multivariate modeling in finance. *International Journal of Managerial Finance, 1*(2), 95-106. doi: 10.1108/17439130510600811

Michaud, R. O. (1989). The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal, 45*(1), 31-42. doi: 10.2469/faj.v45.n1.31

Pachamanova, D. A., Kolm, P. N., Fabozzi, J. F., & Focardi, F. M. (2007). Robust portfolio optimization. In F. J. Fabozzi (Ed.), *Encyclopedia of Financial Models* (Vol. III) (pp. 137-147). Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/9781118182635.efm0095

Perret-Gentil, C., & Victoria-Feser, M.-P. (2004). *Robust mean variance portfolio selection* (Working paper 173). Zürich, Switzerland: National Centre of Competence in Research.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*. 871-880. doi: 10.2307/2288718

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223. doi: 10.2307/1270566

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Hardle, & R. D. Martin (Eds.), *Robust and nonlinear time series analysis* (pp. 256-272). New York, NY: Springer-Verlag. doi: 10.1007/978-1-4615-7821-5_15

Santos, A. A. P. (2010). The out-of-sample performance of robust portfolio optimization. *Revista Brasileira de Finanças, 8*(2), 141-166.

Tütüncü, R., & Koenig, M. (2004). Robust asset allocation. *Annals of Operations Research, 132*(1), 157-187. doi: 10.1023/b:anor.0000045281.41041.ed

Welsch, R. Y., & Zhou, X. (2007). Application of robust statistics to asset allocation models. *REVSTAT – Statistical Journal, 5*(1), 97-114. Retrieved from https://ine.pt/revstat/pdf/rs070106.pdf

Werner, M. (2003). *Identification of multivariate outliers in large data sets* (Doctoral dissertation). University of Colorado Denver, Denver, CO. Retrieved from http://math.ucdenver.edu/graduate/thesis/werner_thesis.pdf

Zhou, X. (2006). *Application of robust statistics to asset allocation models* (Master's thesis). Massachusetts Institute of Technology, Cambridge, MA. Retrieved from http://hdl.handle.net/1721.1/36231

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 612.

ISSN 1538 – 9472

Algorithms and Code

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 613-629.
doi: 10.22237/jmasm/1493598780

ISSN 1538 − 9472

JMASM43: TEEReg: Trimmed Elemental Estimation (R)

Wei Jiang
University of Kansas Medical Center
Kansas City, KS

Matthew S. Mayo
University of Kansas Medical Center
Kansas City, KS

Trimmed elemental regression is robust to outliers and violations of model assumptions. Its properties and statistical inference were evaluated using bias-corrected and accelerated bootstrap confidence intervals. An R package named TEEReg is developed to compute the trimmed elemental estimates and the corresponding bootstrap confidence intervals. Two examples are provided to demonstrate its usage.

Keywords: Trimmed elemental estimator, robust linear regression, R, bias-corrected and accelerated bootstrap confidence interval

Introduction

Linear regression is useful in discovering relationships between observations and covariates. Assume that $\mathbf{Y}$ is an n-dimensional vector of dependent variables, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters, $\boldsymbol{\epsilon}$ is an n-dimensional vector of random errors with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, and $\mathbf{X}$ is a design matrix with n rows and p columns, the multiple linear regression model can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

For the ordinary least square (OLS) approach, the estimator

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

minimizes the sum of squares of the residuals

Dr. Jiang is a Ph.D candidate. Email them at: willjiang29@gmail.com. Dr. Mayo is a professor in the Department of Biostatistics.

$$\hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}} = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^t \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

Although the OLS approach has advantages of easy calculation and well-developed statistical inference, it is sensitive to outliers and violations of model assumptions.

The weighted least square (WLS) and iterative reweighted least square (IRLS) are commonly employed alternatives to the OLS approach to deal with unequal variances of the error terms and influential outlying observations; see Kutner, Nachtsheim, Neter, and Li (2005) for a complete review. Other examples of IRLS can be found in Schlossmacher (1973), Sposito, Kennedy, and Gentle (1977), Krasker and Welsch (1983), Carroll and Ruppert (1988), and Street, Carroll, and Ruppert (1988). There are some other available alternatives to OLS. In 1760, Boscovich first introduced the absolute values estimator that was put into a more structured form later by Laplace (Dielman, 2005). The concept of regression quantiles was generalized by Koenker and Bassett (1978); see also Koenker and D'Orey (1987), Gutenbrunner and Jureckova (1992), Koenker (1994), and Koenker (2005). The least median of squares regression was developed by Rousseeuw (1984), and Hawkins (1993) introduced the globally best estimator and the best elemental estimator. Most of these alternatives were developed based on modifying fitting criteria.

The trimmed elemental (TE) estimator that is robust to outliers and violations of model assumptions was developed by Mayo and Gray (1997). It belongs to a class of regression estimators called leverage-residual weighted elemental (LRWE) estimators (Mayo & Gray, 1997). Hall and Mayo (2008) explored the inference properties of TE approach by investigating the coverage probability of the associated bias-corrected and accelerated (BCa) bootstrap confidence interval (CI). Compared with the traditional bootstrap methods, the BCa approach proposed by Efron (1987) corrected the bias and skewness of the sampling distribution through adjusting the selected percentiles used for constructing CIs.

The purpose of this article is to provide an R-package called TEEReg to compute the TE estimates and the corresponding BCa bootstrap CIs. This package contains two functions, TEE() and TEE.BCa(), and can be obtained at CRAN at http://cran.r-project.org/web/packages/TEEReg/.

TE Estimator and BCa Bootstrap CI

The TE estimator developed by Mayo and Gray (1997) is robust to outlying cases and violations of model assumptions. It is a solution based on the elemental subset and the elemental regressions.

Elemental Subsets and Elemental Regressions

In most situations, the sample size n is much larger than the number of unknown parameters p. Instead of using all n observations, only p are required to obtain estimates of the p-dimensional vector of unknown parameters defined in model (1). In this case, there are $\binom{n}{p}$ distinct subvectors of the data and thus $\binom{n}{p}$ possible solutions for the vector $\boldsymbol{\beta}$ in which each solution provides an exact fit to the corresponding p observations. Let $h = \{i_1, i_2, \ldots, i_p\}$ be a subset containing p distinct values from the n-dimensional set of indices $\{1, 2, \ldots, n\}$, $\mathbf{X}_h$ denote a p-dimensional square matrix constructed by the rows of $\mathbf{X}$ with corresponding indices, and $\mathbf{Y}_h$ denote a $p \times 1$ subvector of $\mathbf{Y}$ of which elements are those in $\mathbf{Y}$ indexed by the subset h. Then, the subset h is an elemental subset of the data and the solution to $\mathbf{X}_h\hat{\boldsymbol{\beta}}_h = \mathbf{Y}_h$, a system of p equations with p unknowns, is called an elemental regression and is given by

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \frac{\sum_h \left|\mathbf{X}_h^t\mathbf{X}_h\right|\hat{\boldsymbol{\beta}}_h}{\sum_h \left|\mathbf{X}_h^t\mathbf{X}_h\right|} = \sum_h \frac{\left|\mathbf{X}_h^t\mathbf{X}_h\right|}{\left|\mathbf{X}^t\mathbf{X}\right|}\hat{\boldsymbol{\beta}}_h = \sum_h w_h\hat{\boldsymbol{\beta}}_h \tag{2}$$

where $|\mathbf{A}|$ denotes the determinant of matrix $\mathbf{A}$. This indicates that the least squares estimate is a weighted average over all possible elemental estimates $\hat{\boldsymbol{\beta}}_h$ with weights

$$w_h = \frac{\left|\mathbf{X}_h^t\mathbf{X}_h\right|}{\left|\mathbf{X}^t\mathbf{X}\right|}$$

Moreover, Mayo and Gray (1997) demonstrated that the WLS estimator can be formed as a function of elemental regressions. Let v_i denote the weight for observation i, $\mathbf{V}$ be a diagonal matrix containing the weights v_i, and $\mathbf{V}_h$ be a $p \times p$

submatrix of **V** corresponding to the elemental subset *h*. After some calculations, the WLS estimator can be equivalently written as

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = \frac{\sum_h \left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right| \hat{\boldsymbol{\beta}}_h}{\sum_h \left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right|} = \sum_h \frac{\left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right|}{\left|\mathbf{X}^t \mathbf{V} \mathbf{X}\right|} \hat{\boldsymbol{\beta}}_h = \sum_h w_h^* \hat{\boldsymbol{\beta}}_h \tag{3}$$

In practice, the reciprocal of the variances of error terms is usually employed for weight v_i to deal with unequal error variances (Kutner et al., 2005), so a lesser weight is assigned to an observation with a larger variance than another observation with a smaller variance. Many weight functions were suggested for dampening the influence of outlying observations, including the Huber weight function given below (Kutner et al., 2005):

$$v_i = \begin{cases} 1 & |u_i| \le 1.345 \\ \dfrac{1.345}{|u_i|} & |u_i| > 1.345 \end{cases}$$

where u_i denotes the scaled residual for which a definition can be found in Kutner et al. (2005). It does not reduce the weight of a case from 1 until the absolute scaled residual is greater than 1.345. It is usually suggested to re-estimate the scaled residual using the process of IRLS to obtain revised weights when the initial estimated coefficients are substantially different from the ones obtained by OLS (Kutner et al., 2005).

TE Estimator

The TE estimator is a special case of a class of estimators called leverage-residual weighted elemental (LRWE) estimators developed by Mayo and Gray (1997). The LRWE class consists of all estimators that can be expressed in the form

$$\hat{\boldsymbol{\beta}}(\lambda, \rho) = \frac{\sum_h w\left[\lambda(h), \rho(h)\right] \hat{\boldsymbol{\beta}}_h}{\sum_h w\left[\lambda(h), \rho(h)\right]}$$

where the factor $\lambda(h)$ represents the leverage information related to the elemental subsets *h* and the factor $\rho(h)$ represents the information of degree of fit related to elemental subsets. The OLS estimator defined in formula (2) belongs to the class

of LRWE estimators with $\lambda(h) = \left|\mathbf{X}_h^t \mathbf{X}_h\right|$, $\rho(h) = 1$, and $w[\lambda(h), \rho(h)] = \lambda(h)\rho(h)$. This reveals that the OLS approach only considers the information of leverage but does not take the information of degree of fit for each elemental subset h into account; the resulting estimates can be easily affected by the influential points. Moreover, it can be seen from formula (3) that the WLS estimator is a member of the LRWE class with $\lambda(h) = \left|\mathbf{X}_h^t \mathbf{X}_h\right|$, $\rho(h) = |\mathbf{V}_h|$, and $w[\lambda(h), \rho(h)] = \lambda(h)\rho(h)$. This is because $\mathbf{X}_h$ and $\mathbf{V}_h$ are square matrices and $\left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right| = \left|\mathbf{X}_h^t \mathbf{X}_h\right|\left|\mathbf{V}_h\right|$. This explains why the WLS approach is robust to violations of model assumptions and influential observations because it considers the information of both leverage and degree of fit.

Mayo and Gray (1997) developed a robust TE estimator based on the LRWE class. Unlike the OLS method where the same weight of degree of fit is assigned to all elemental regressions regardless of whether they are influenced by outlying cases, the TE method removes or trims out those elemental regressions that poorly fit the data due to extreme observations from calculations. With $\lambda(h)$ and $\omega[\lambda(h), \rho(h)]$ remaining the same as those in formula (2), the TE estimator alters $\rho(h)$ to have the form

$$\rho(h) = \begin{cases} 1, & \text{if } rank\left(\sum_{i=1}^{n} |e_{hi}|\right) \le (1-\alpha_p)\binom{n}{p} \\ 0, & \text{otherwise} \end{cases}$$

where α_p represents the trimming proportion that ranges from 0 to 1 and $\sum_{i=1}^{n} |e_{hi}|$ is the sum of absolute residuals based on the elemental estimates $\hat{\boldsymbol{\beta}}_h$. By ruling out those elemental regressions adversely affected by extreme cases, the TE approach produces estimators robust to outliers and violations of model assumptions. Notice that the degree of robustness of the presented approach depends on the values selected for trimming proportion α_p. A bigger α_p means a greater robustness because it removes more elemental regressions with large sums of absolute residuals than a lower α_p does. Depending on the proportion of regressions one would like to remove from consideration, α_p can be adjusted accordingly. Taking this into account, the TE estimator is denoted as TEE(α_p).

The TE approach is different from eliminating outliers from data. Omission of outlying observations takes away multiple elemental subsets including some good ones that could potentially exist with those observations. For example, if a

dataset contains 10 observations and 2 unknown parameters are of interest, there are $\binom{10}{2} = 45$ elemental regressions total. If one outlier is removed, then the total number of elemental regressions reduces to $\binom{9}{2} = 36$. As you may expect, the number of elemental regressions eliminated from analysis increases dramatically as *n* or *p* becomes bigger. Deleting observations from data is not the best way to handle outliers unless the outlying cases are indeed resulted from mistakes or other extraneous causes.

BCa Bootstrap CI

The BCa approach, suggested by Efron (1987), seeks to correct the bias and skewness of the sampling distribution through adjusting the selected percentiles used for constructing CIs. The adjusted percentiles are

$$\delta_1 = \phi\left(\hat{z} + \frac{\hat{z} + z_{\alpha/2}}{1 - \hat{\alpha}\left(\hat{z} + z_{\alpha/2}\right)}\right) \quad \text{and} \quad \delta_2 = \phi\left(\hat{z} + \frac{\hat{z} + z_{1-\alpha/2}}{1 - \hat{\alpha}\left(\hat{z} + z_{1-\alpha/2}\right)}\right)$$

where $\phi(.)$ is the standard normal cumulative function and z_α represents the $100\alpha\%$ quantile of the standard normal distribution. The skewness and bias of the sampling distribution are respectively adjusted by $\hat{z}$ and $\hat{\alpha}$, expressions of which can be found in Efron (1987) and DiCiccio and Efron (1996). In general, the algorithm for creating the $100(1-\alpha)\%$ BCa bootstrap CIs in terms of the TE estimation is given as follows:

- For $m = 1,\ldots, M$, do:
  - (a) Sample data with replacement from the dataset.
  - (b) Compute TE estimates $\hat{\boldsymbol{\beta}}_{\text{TEE}}$ based on the m^{th} bootstrap sample.
- Construct the $100(1-\alpha)\%$ BCa bootstrap CIs using the adjusted percentiles given above based on the generated bootstrap sample of $\hat{\boldsymbol{\beta}}_{\text{TEE}}$

Hall and Mayo (2008) conducted simulation studies under various scenarios to compare the coverage probabilities of BCa bootstrap CIs based on the TE estimation to the ones based on other approaches. It was found that the BCa bootstrap CIs in terms of TE estimators are almost indistinguishable from those based on OLS when error terms follow the Normal, Contaminated Normal, or

Student's *t* distribution. For the Cauchy and Laplace error distributions, however, the TE estimation is preferred (See Hall and Mayo (2008) for more details). This indicates the OLS estimator is robust to small departures from normality; however, major departures from normality should be of concern.

Computation Efficiency

Even with powerful computers available today the computation time for deriving TE estimates increases tremendously as the number of regression parameters or sample size increases. For example, if there are 10 observations and the model only has two parameters, then $\binom{10}{2} = 45$ elemental subsets need to be fit; however, if the sample size and number of parameters increase to 20 and 4, respectively, we need to fit $\binom{20}{4} = 4845$ elemental regressions, which requires over 100 times more computations. In order to reduce the computation intensity, Hall and Mayo (2008) examined the appropriateness of the approach of random subsample, suggested by Hawkins (1993) for the best elemental estimator, for reducing the number of computations required for the TE estimator through simulation studies. They claimed that computing the TE estimates based on as low as 50% of the elemental subsets may be sufficient to produce reliable estimates as long as the error terms follow Normal, Cauchy, Laplace, 10% Contaminated Normal, or Student's *t* distribution.

TEEReg Package

The proposed R package TEEReg provides tools for computing the TE estimates and the corresponding BCa bootstrap CIs. In this section, the usage of the two functions TEE() and TEE.BCa() enclosed in TEEReg are explained.

The function TEE() is used to compute the TE estimates. Its usage with complete arguments is given as:

```
TEE(formula, data, offset=NULL, p.trimmed=NULL, p.subsample=1,
method="tee")
```

Similar to other R functions developed for linear regressions, such as lm() and glm(), the first argument formula gives a symbolic description of the model to be fitted (e.g. formula = $y \sim x$). The second argument specifies the dataset used

for performing regression analyses. Be aware that the data must be formatted as a data frame prior to using the TEE() function. The offset can be used to specify regressors with coefficients of 1. This argument can be either NULL or a numeric vector with length equal to the number of observations. The argument p.trimmed indicates the proportion of elemental subsets removed from the computation of estimates. It should be either NULL or a numeric value between 0 and 1. However, a value must be provided to p.trimmed when method = "tee" is specified. The argument p.subsample is for specifying the proportion of random selection of elemental subsets. One may improve the computation efficiency by providing a numeric value between 0 and 1 to this argument. The default value of p.subsample is 1 under which the TE estimates are calculated based on all elemental subsets. When using the TEE() function, the TE regression is carried out by default (i.e., the default value to argument method is "tee"). Another supported option for this argument is "ols" under which the OLS approach is employed for fitting linear regressions. When the value ols is given to the argument of method, the TEE() function computes the estimates based on the full data no matter what values are assigned to p.trimmed and p.subsample.

The second function TEE.BCa() is used to construct the $100(1-\alpha)\%$ BCa bootstrap CIs based on the TE estimation. It is similar in structure to TEE() and has the form with complete arguments as follows:

```
TEE.BCa(formula, data, offset=NULL, p.trimmed=NULL, p.subsample=1,
method="tee", est.TEE, conf.level, n.boot)
```

The specifications of the first six arguments in TEE.BCa() are the same as explained above for TEE(). For the remaining three, est.TEE stands for TE regression estimates, and conf.level and n.boot represent the confidence level and the number of bootstrap samples, respectively. Detailed descriptions of the arguments enclosed in these two functions can also be viewed using the command ??TEE.

Sometimes, the elemental regression $\hat{\boldsymbol{\beta}}_h$ is not estimable because $\mathbf{X}_h$ is singular and the inverse matrix $\mathbf{X}_h^{-1}$ does not exist. This could happen, for example, when several subjects have the same covariates values and so the matrix $\mathbf{X}_\mathrm{h}$ is not full-rank. The TEEReg package handles such situations using the Moore-Penrose generalized inverse, which is defined and unique for all matrices whose entries are real or complex numbers. It is computed using the singular

value decomposition. For a review of the Moore-Penrose generalized inverse, see Campbell and Meyer (2009).

Examples

To evaluate the robustness of the presented TE approach, the first example is based on the telephone data (Rousseeuw & Leroy, 1987) with several outlying observations and the second example is simulated data based on a Cauchy distribution. For both examples, the 95% BCa bootstrap CIs are created based on 1000 bootstrap samples.

Example 1: Data with Outliers

In this example, the telephone data (Rousseeuw & Leroy, 1987) are used to demonstrate the usage of the TEEReg package. In the data, the number of telephone calls (tens of millions) made in Belgium was recorded from 1959 to 1973. It contains several extreme observations resulted from mistakes in recording units over the years 1964-1969 (see Figure 1), which is useful in order to examine the robustness of the TE method to outliers. The response variable of the telephone data is the number of telephone calls and the independent variable is the year. For illustration purposes, the TE estimates and the corresponding 95% BCa bootstrap CIs are computed based on both 30% and 42% trimming proportions. The results in terms of all elemental subsets and those based on 70% random subsample are also compared in this example.

The TEEReg package can be loaded into R by the command library(TEEReg). The telephone data are stored inside the package and can be accessed by the command data(telephone). As explained above, the TE estimates and the corresponding 95% BCa bootstrap CIs in terms of the subsample proportion of 100% and trimming proportion of 42% can be computed by typing the following:

```
R> fitTEE1 <- TEE(formula=Y~X, data=telephone, p.trimmed=0.42,
p.subsample=1, method="tee")
R> CITEE1 <- TEE.BCa(formula=Y~X, data=telephone, p.trimmed=0.42,
p.subsample=1, + method="tee", est.TEE=fitTEE1$coefficients,
conf.level=0.05, n.boot=1000)
```

Their outputs are displayed as below:

```
R> fitTEE1
$call
TEE(formula = Y ~ X, data = telephone, p.trimmed = 0.42, p.subsample = 1,
method = "tee")
$formula
Y ~ X
$coefficients
(Intercept)        X
-100.0543   1.991974
$residuals
       1          2          3          4         5          6          7
 4.855597   3.163623   1.171649  0.3796743   -0.9123  -2.204274  -3.396248
       8          9         10         11        12         13         14
-4.688223  -4.880197  -5.472171  -5.964145  -6.55612  -7.348094  -4.240068
      15         16         17         18        19         20         21
 91.56796   94.57598    110.584    125.592  146.6001   174.6081   3.616112
      22         23         24
-17.37586  -16.36784  -16.35981
$fitted.values
       1         2          3         4         5         6         7
-0.455597  1.536377   3.528351  5.520326    7.5123  9.504274  11.49625
       8         9         10        11        12        13        14
 13.48822   15.4802   17.47217  19.46415  21.45612  23.44809  25.44007
      15        16         17        18        19        20        21
 27.43204  29.42402   31.41599  33.40797  35.39994  37.39191  39.38389
      22        23        24
 41.37586  43.36784  45.35981

R> CITEE1
$call
TEE.BCa(formula = Y ~ X, data = telephone, p.trimmed = 0.42, p.subsample = 1,
method = "tee", est.TEE = fitTEE1$coefficients, conf.level = 0.05, n.boot =
1000)
$ci
             estimates(TEE)  Lower limit  Upper limit
(Intercept)       -100.0543  -452.481442   -49.220453
X                  1.991974     1.045627     8.588198
```

Note the output yielded by the function TEE() contains the model formula, estimates of coefficients, residuals, and fitted values, and the output of the TEE.BCa() function consists of the model formula and BCa bootstrap CIs for regression parameters. In the case that one only wants to extract, for example, the coefficient estimates from the output of TEE() function, the command fit1$coefficients can be used. The TE estimates and the corresponding 95% BCa bootstrap CIs based on other scenarios planned to be investigated in this example can be computed following a similar manner by specifying p.trimmed = 0.30 and p.subsample = 1 or 0.7. The key results are summarized in Table 1. For comparison purposes, the results based on the OLS approach and the IRLS using Huber weight function are also presented in this table.

The estimated regression function using the TE approach with p.subsample = 1 and 42% trimming suggests that the mean number of telephone calls are expected to increase by 1.992 (in tens of millions) when the year increases by 1. The corresponding 95% BCa bootstrap CI for the slope is (1.046, 8.588) which does not include 0. Based on this scenario, it can be concluded that year is significantly linearly related to the number of telephone calls. As expected, the outlying observations are more influential in the fitted TE regression function with p.subsample = 1 and 30% trimming proportion. The estimated slope is dragged up by outliers to 3.940 (BCa CI: 1.114, 8.424) due to the fact that more elemental regressions with large sums of absolute residuals are used in calculations. The same trend can be observed in the case of p.subsample = 0.7.

Moreover, it can be seen in Table 1 that the TE estimates based on 70% random subsample of elemental subsets are similar to those based on all elemental subsets for both cases of TEE(30%) and TEE(42%). The 95% BCa bootstrap CIs in terms of 70% subsample are wider than the ones based on all elemental subsets, but both lead to the same conclusion of statistical inference. It seems that using the 70% subsampling provides fairly accurate estimates and works almost equally well as utilizing the full data for the given telephone data.

Table 1. Estimates of coefficients and 95% BCa bootstrap Cis based on various approaches using telephone data

Methods	Intercept est.	95% CI (intercept)	Slope est.	95% CI (slope)
TEE(30%): p.subsample = 1	-204.034	(-452.688, -52.983)	3.940	(1.114, 8.424)
TEE(30%): p.subsample = 0.7	-217.143	(-516.187, -54.649)	4.193	(1.145, 9.520)
TEE(42%): p.subsample = 1	-100.054	(-452.481, -49.220)	1.992	(1.046, 8.588)
TEE(42%): p.subsample = 0.7	-112.678	(-540.452, -50.289)	2.235	(1.062, 10.069)
OLS	-260.059	(-523.136, -118.906)	5.041	(2.475, 9.549)
IRLS	-99.904	(-590.294, -52.987)	1.987	(1.113, 10.873)

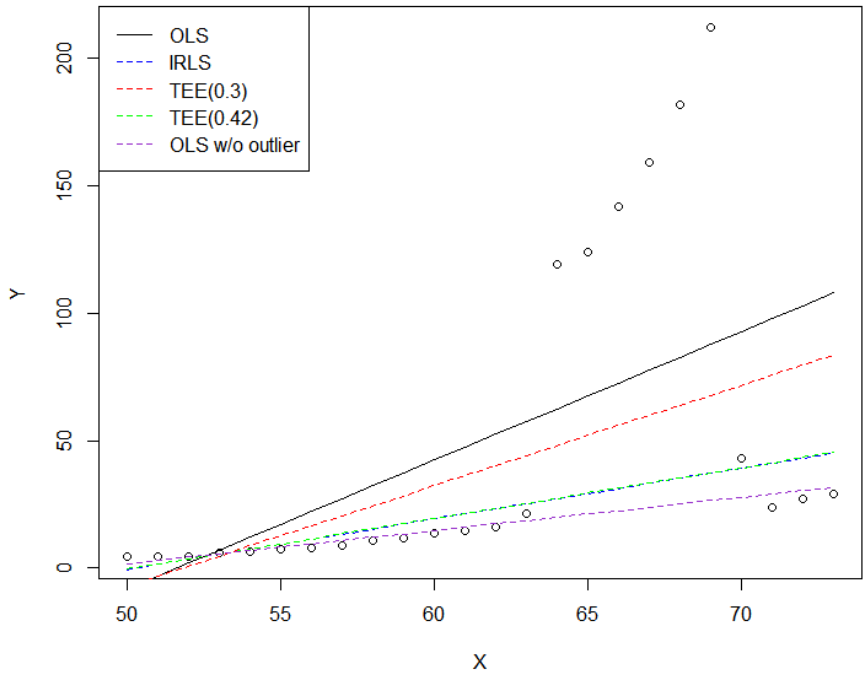


Figure 1. Fitted regression lines using different regression approaches for telephone data

Figure 1 displays the fitted regression lines for a variety of regression approaches. The overlaid TE regression lines are obtained in terms of all elemental subsets (i.e., p.subsample = 1). In addition, a regression line fitted using the OLS approach based on the telephone data with outliers removed is also included in this figure for comparison purposes. It is obvious that the OLS approach performs the worst with its estimates dramatically affected by outliers. The regression lines based on IRLS and TEE(42%) are overlapped with each other because they lead to almost identical estimates of unknown parameters (see Table 1). This is not surprising because the IRLS approach is also robust to outlying cases. The 95% BCa bootstrap CIs for IRLS are wider than the ones for TEE(42%) (see Table 1). As explained in the previous paragraph, due to the fact that relatively more elemental regressions having large sums of absolute residuals are employed in calculations, the TEE(30%) is affected more by the outliers than the TEE(42%) and IRLS. Both fitted regression lines of TEE(30%) and TEE(42%) are above the one based on the OLS approach with outliers removed. The reason is that deleting outlying observations takes away all of their corresponding elemental subsets.

Example 2: Cauchy Data

In this example, a simulated dataset consisting of 50 observations and one independent variable is used to clarify the usage of TEEReg package and to illustrate the robustness to non-normal data of the presented TE estimator. The values of the independent variable X are generated from a Poisson distribution with mean equal to 10 and the values of the dependent variable Y are computed as $Y = 0.5 + 1X + \epsilon$, where the error term ϵ is assumed to follow a Cauchy distribution with location 0 and scale 1. We call this artificial dataset the data.sim. In this example, the TE estimates and the corresponding 95% BCa bootstrap CIs are computed based on all elemental subsets and both 50% and 75% trimming proportions. As demonstrated in Hall and Mayo (2008), these two trimming proportions provide high coverage probabilities (at least 95%) to the 95% BCa bootstrap CIs when the error term follows Cauchy distribution.

The TE estimates and the corresponding 95% BCa bootstrap CIs in terms of the subsample proportion of 100% and trimming proportion of 50% can be computed by typing the following:

```
R> fitTEE3 <- TEE(formula=Y~X, data=data.sim, p.trimmed=0.5,
p.subsample=1,method = "tee")
R> CITEE3 <- TEE.BCa(formula=Y~X, data=data.sim, p.trimmed=0.5,
p.subsample=1, + method="tee", est.TEE=fitTEE3$coefficients,
conf.level=0.05, n.boot=1000)
```

The TE estimates and their BCa CIs based on 75% trimming can be computed similarly by specifying p.trimmed = 0.75. The key outputs of both scenarios are summarized in Table 2. For comparison purposes, the results based on the OLS method and the IRLS using Huber weight function are also given in this table.

Table 2. Estimates of coefficients and 95% BCa bootstrap Cis based on various regression approaches using simulated data

Methods	Intercept est.	95% CI (intercept)	Slope est.	95% CI (slope)
TEE(50%)	1.341	(-0.542 , 6.602)	0.899	(0.305, 1.120)
TEE(75%)	0.919	(-1.026, 3.734)	0.967	(0.634, 1.170)
OLS	6.639	(1.858, 12.516)	0.471	(-0.096, 0.934)
IRLS	2.100	(0.0728, 7.281)	0.832	(0.240, 1.055)

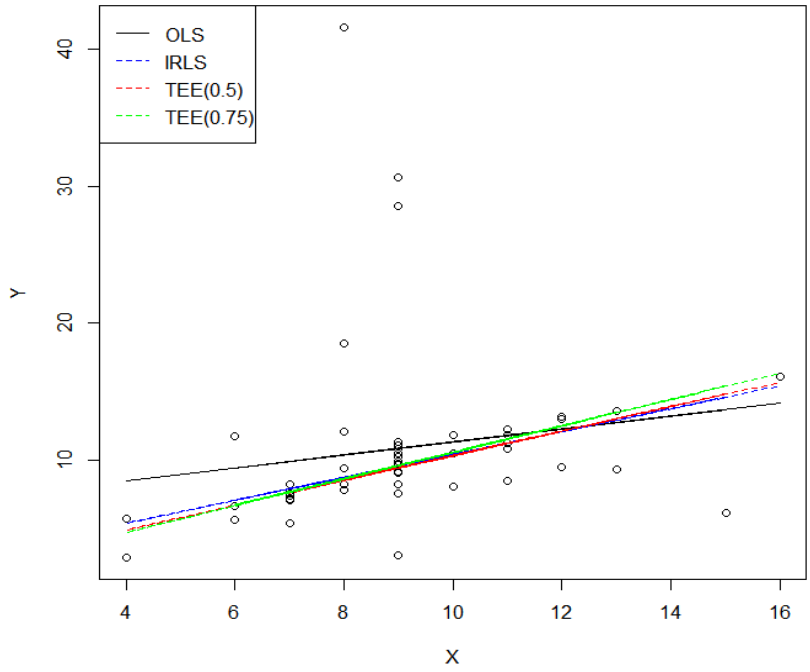


Figure 2. Fitted regression lines using different regression approaches for simulated data

As expected, the OLS approach performs the worst in terms of handling the simulated Cauchy data. The corresponding 95% BCa bootstrap CIs for intercept and slope are, respectively, (1.858, 12.516) and (-0.096, 0.934), none of which captures the true values of 0.5 and 1. The OLS estimates of both intercept and slope are significantly different from the true values as well. In contrast, it appears that the TEE(75%) performs the best for the given dataset. The resulting TE estimates for slope and intercept are, respectively, 0.919 and 0.967, both of which are very close to the true intercept and slope used for generating data. The estimates produced by TEE(50%) seems to be slightly worse than ones based on TEE(75%), but it is closer to the true values than the ones resulting from IRLS. The 95% BCa bootstrap CIs of both TEE(50%) and IRLS contain the true intercept and slope of 0.5 and 1. It appears that the TE approach is robust to the simulated Cauchy data that severely depart from normality. A scatterplot of the simulated data along with fitted regression lines using different approaches is shown in Figure 2.

Summary

The usage of a new R package TEEReg was explicated for computing the TE estimates and creating the BCa bootstrap CIs. This package includes two functions: TEE() for the TE regression and TEE.BCa() for the BCa bootstrap CIs. Two examples were provided in this paper to demonstrate the usage of the TEEReg package. In the first example, the telephone data with several influential observations were used to examine the robustness of the TE method to outliers. It was found that the TEE(42%) and IRLS approaches work equally well for the given dataset. The TEE(30%) was affected more by the outliers because, compared to $\alpha_p = 42\%$, relatively more elemental regressions with large sums of absolute residuals are involved in calculations. The random subsample approach, suggested by Hawkins (1993), was employed in this example as well. It appeared that, for the telephone dataset, using the 70% subsampling provides fairly accurate estimates and works almost equally well as utilizing the full data. This is consistent with the conclusions of Hall and Mayo (2008), that the random subsample approach is appropriate for reducing computation intensity when the error terms follow certain distributions. In the second example, a simulated data set with Cauchy error terms was used to assess the robustness of the TE approach to non-normal data. It appeared that the TE estimator is robust and efficient to the simulated data with Cauchy error terms. This is also consistent with the findings based on simulation studies from Hall and Mayo (2008). The new TEEReg package can be readily used to conduct TE regression analysis which is a useful and robust alternative to OLS in the presence of outliers and violations of model assumptions.

References

Campbell, S. L., & Meyer, C. D. (2009). *Generalized inverses of linear transformations*. Philadelphia, PA: Society for Industrial and Applied Mathematics. doi: 10.1137/1.9780898719048

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York, NY: Chapman and Hall.

DiCiccio, T., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3), 189-212. Available from: http://www.jstor.org/stable/2246110

Dielman, T. E. (2005). Least absolute value regression: Recent contributions. *Journal of Statistical Computation and Simulation, 75*(4), 263-286. doi: 10.1080/00949650420002236 80

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171-185. doi: 10.2307/2289144

Gutenbrunner, C., & Jureckova, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics, 20*(1), 305-330. doi: 10.1214/aos/1176348524

Hall, M., & Mayo, M. S. (2008). Bootstrap confidence intervals and coverage probabilities of regression parameter estimates using trimmed elemental estimation. *Journal of Modern Applied Statistical Methods, 7*(2), 514-525. Retrieved from: http://digitalcommons.wayne.edu/jmasm/vol7/iss2/17/

Hawkins, D. M. (1993). The accuracy of elemental set approximations for regression. *Journal of the American Statistical Association, 88*(422), 580-589. doi: 10.2307/2290339

Koenker, R. W. (1994). Confidence intervals for regression quantiles. In P. Mandl, M. Hušková (Eds.), *Asymptotic statistics* (pp. 349-359). New York, NY: Springer-Verlag. doi: 10.1007/978-3-642-57984-4_29

Koenker, R. W. (2005). *Quantile regression*. New York, NY: Cambridge University Press. doi: 10.1017/cbo9780511754098

Koenker, R., & Bassett, G. J. (1978). Regression quantiles. *Econometrica, 46*(1), 33-50. doi: 10.2307/1913643

Koenker, R. W., & D'Orey, V. (1987). Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 36*(3), 383-393. doi: 10.2307/2347802

Krasker, W. S., & Welsch, R. E. (1983). The use of bounded-influence regression in data analysis: theory, computation, and graphics. *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*. New York, NY: Springer-Verlag.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw-Hill.

Mayo, M. S., & Gray, J. B. (1997). Elemental subsets: The building blocks of regression. *The American Statistician, 51*(2), 122-129. doi: 10.2307/2685402

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871-880. doi: 10.2307/2288718

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: John Wiley and Sons. doi: 10.1002/0471725382

Schlossmacher, E. J. (1973). An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association, 68*(344), 857-859. doi: 10.2307/2284512

Sposito, V. A., Kennedy, W. J., & Gentle, J. E. (1977). Algorithm AS 110: L_p norm fit of a straight line. *Journal of the Royal Statistical Society*. *Series C (Applied Statistics), 26*(1), 114-118. doi: 10.2307/2346888

Street, J. O., Carroll, R. J., & Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician, 42*(2), 152-154. doi: 10.2307/2684491

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 630-656.
doi: 10.22237/jmasm/1493598840

ISSN 1538 − 9472

Multiple Ratio Imputation by the EMB Algorithm: Theory and Simulation

Masayoshi Takahashi
Tokyo University of Foreign Studies
Tokyo, Japan

Although multiple imputation is the gold standard of treating missing data, single ratio imputation is often used in practice. Based on Monte Carlo simulation, the Expectation-Maximization with Bootstrapping (EMB) algorithm to create multiple ratio imputation is used to fill in the gap between theory and practice.

Keywords: Multiple imputation, ratio imputation, Expectation-Maximization, bootstrap, missing data, incomplete data, nonresponse, estimation uncertainty

Introduction

In survey data, missing values are prevalent. At best, missing data are inefficient because the incomplete dataset does not contain as much information as is expected. At worst, missing data can be biased if non-respondents are systematically different from respondents (Rubin, 1987). The best solution to the missing data problem is to collect the true data, by resending questionnaires or by calling respondents. Nevertheless, there are two problems to this ideal solution. First, data users often have no luxury of collecting more data to take care of missingness. Second, facing a worldwide trend of resource reduction in official statistics, data providers such as national statistical agencies need to make the statistical production as efficient as possible. From these two perspectives for both data users and data providers, parametric imputation models, if used properly, may help to reduce bias and inefficiency due to missing values. In fact, if the missing mechanism is at random (MAR), it has been demonstrated that imputation can ameliorate the problems associated with incomplete data (Little & Rubin, 2002; de Waal et al., 2011).

Masayoshi Takahashi is an Assistant Professor of Institutional Research. Email at mtakahashi@tufs.ac.jp.

Among others, ratio imputation is often used to treat missing values in practice (de Waal et al., 2011; Thompson & Washington, 2012; Office for National Statistics, 2014). When there is an auxiliary variable that is a de facto proxy for the target incomplete variable, ratio imputation is assumed to produce high quality data (Hu et al., 2001). On the other hand, proponents of multiple imputation have long argued that single imputation generally ignores estimation uncertainty by treating imputed values as if they were true values (Rubin, 1987; Schafer, 1997; Little & Rubin, 2002). Multiple imputation is indeed known to be the gold standard of handling missing data (Baraldi & Enders, 2010; Cheema, 2014). In the literature, however, there is no such thing as multiple ratio imputation, leading to a gap between theory and practice. Here, we fill in this gap by proposing a novel application of the Expectation-Maximization with Bootstrapping (EMB) algorithm to ratio imputation, where multiple-imputed values will be created for each missing value.

Therefore, the purpose of this study is to examine the standard single ratio imputation techniques and their limitations, illustrate the mechanism and advantages of multiple ratio imputation, and assess the performance of multiple ratio imputation using 45,000 simulated datasets based on a variety of sample sizes, missing rates, and missingness mechanisms. Also, a review of `MrImputation`, provided in Takahashi (2017), is included.

Notations

D is an $n \times p$ dataset, where n is the number of observations and p is the number of variables. If no data are missing, the distribution of **D** is assumed to be multivariate normal, with the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{D} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let i be an observation index, $i = 1,\ldots,n$. Let j be a variable index, $j = 1,\ldots,p$. Thus, $\mathbf{D} = \{Y_1,\ldots,Y_p\}$, where Y_j is the j^{th} column in **D**, and Y_{-j} is the complement of Y_j. Generally, Y_{-j} refers to all of the columns in **D** except Y_j. Especially, this article deals with a two-variable imputation model; thus, Y_1 is the incomplete variable (target variable for imputation) and Y_2 is the complete variable (auxiliary variable). Thus, $\mathbf{D} = \{Y_{i1}, Y_{i2}\}$.

Also, let **R** be a response indicator matrix, whose dimension is the same as **D**. Whenever **D** is observed **R** = 1, and whenever **D** is not observed **R** = 0. Note, *R* in Italics refers to the *R* software environment for statistical computing and graphics. $\mathbf{D}_{\mathbf{obs}}$ refers to the observed part of data, and $\mathbf{D}_{\mathbf{mis}}$ refers to the missing part of data, i.e., $\mathbf{D} = \{\mathbf{D}_{\mathbf{obs}}, \mathbf{D}_{\mathbf{mis}}\}$. β is the slope in the complete model, $\hat{\beta}$ is the

slope estimated by the observed model, and $\tilde{\beta}$ is the estimated slope by multiple imputation.

Assumptions of Missing Mechanisms

There are three assumptions of missingness (Little & Rubin, 2002; King et al., 2001). This is an important issue, because the results of statistical analyses depend on the type of missing mechanisms (Iwasaki, 2002). The first assumption is Missing Completely At Random (MCAR), which means that the missingness probability of a variable is independent of the data for the unit. In other words, $P(\mathbf{R}|\mathbf{D}) = P(\mathbf{R})$. Take an economic survey where enterprises choose to answer their turnover values by tossing a coin as a perfect example of MCAR. This is the easiest case to take care of, because MCAR is simply a case of random subsampling from the intended sample; thus, subsamples may be inefficient, but unbiased. Note that the assumption of MCAR can be tested by entering dummy variables for each variable, and scoring it 1 if the data are missing and 0 otherwise.

The second assumption is the case where missingness is conditionally at random. Traditionally, this is known as Missing At Random (MAR), which means that the conditional probability of missingness given data is equal to the conditional probability of missingness given observed data. In other words, $P(\mathbf{R}|\mathbf{D}) = P(\mathbf{R}|\mathbf{D}_{obs})$. An example of MAR would be when enterprises with few employees, in the above hypothetical survey, are found more likely to refuse to answer their turnover values, assuming that there is a column in the dataset that has values on the number of employees. If the missing mechanism is at random, imputation can rectify the bias due to missingness. Note that the assumption of MAR (unlike MCAR) cannot be tested.

The third assumption is Non-Ignorable (NI), where the missingness probability of a variable depends on the variable's value itself, and this relationship cannot be broken conditional on observed data. In other words, $P(\mathbf{R}|\mathbf{D}) \neq P(\mathbf{R}|\mathbf{D}_{obs})$. Imagine that enterprises with lower values of turnover are more likely to refuse to answer their turnover values in our survey, and the other variables in the dataset cannot be used to predict which enterprises have small amounts of turnover: this would be an example of NI. If the missing mechanism is NI, a general-purpose imputation method may not be appropriate. Instead, a special technique should be developed to take care of the unique nature of non-ignorable missing mechanisms.

For the missingness mechanism to be ignorable, both of the MAR and distinctness conditions need to be met (Little & Rubin, 2002, pp.119-120).

However, under practical conditions, the missingness data model is often regarded as ignorable if the MAR condition is satisfied (Allison, 2002, p.5; van Buuren, 2012, p.33). This means that NI is Not Missing At Random (NMAR).

Also, as Carpenter & Kenward (2013) noted, MAR means that the probability of observing a variable's value often depends on its own value, but the dependence can be eliminated, given observed data. NI means that the probability of observing a variable's value not only depends on its own value, but also the dependence cannot be eliminated, given observed data. However, the meaning of MAR differs from researcher to researcher (Seaman et al., 2013); thus, there is some ambivalence to this terminology.

Existing Algorithms and Software for Multiple Imputation

There are three major algorithms for multiple imputation. The first traditional algorithm is based on Markov chain Monte Carlo (MCMC). This is the original version of Rubin's (1978, 1987) multiple imputation. *R*-Package `Norm` currently implements this version of multiple imputation (Schafer, 1997; Fox, 2015). A commercial software program using the MCMC algorithm is SAS `Proc MI` (SAS, 2011). The second major algorithm is called Fully Conditional Specification (FCS), also known as chained equations by van Buuren (2012). *R*-Package `MICE` currently implements this version of multiple imputation (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren & Groothuis-Oudshoorn, 2015). Other commercial software programs using the FCS algorithm are SPSS `Missing Values` (SPSS, 2009) and SOLAS (Statistical Solutions, 2011). The FCS algorithm is known to be flexible. The third relatively new algorithm is the Expectation-Maximization with Bootstrapping (EMB) algorithm by Honaker & King (2010). *R*-Package `Amelia II` currently implements this version of multiple imputation (Honaker et al., 2011; Honaker et al., 2015). The EMB algorithm is known to be computationally efficient.

Assessing superiority among the different multiple imputation algorithms is beyond the scope of the current study. According to Takahashi & Ito (2013), if the underlying distribution can be approximated by a multivariate normal distribution with the MAR condition, all of the three algorithms essentially give the same answers. As for the performance of the EMB algorithm, Honaker & King (2010) contended the estimates of population parameters in bootstrap resamples can be appropriately used instead of random draws from the posterior. Rubin (1987) argued the approximately Bayesian bootstrap method is proper imputation because it incorporates between-imputation variability. Also, Little & Rubin

(2002) opined the substitution of Maximum Likelihood Estimates (MLEs) from bootstrap resamples is proper because the MLEs from the bootstrap resamples are asymptotically identical to a sample drawn from the posterior distribution. Therefore, multiple imputation by the EMB algorithm can be considered to be proper imputation in Rubin's sense (1987). Also, according to van Buuren (2012), the bootstrap method is computationally efficient because there is no need to make a draw from the χ^2 distribution, unlike the other traditional algorithms of multiple imputation. This means that it is not necessary to resort to the Cholesky decomposition (factorization), the property of which is that if **A** is a symmetric positive definite matrix, i.e., $\mathbf{A} = \mathbf{A}^{\mathbf{T}}$, then there is a matrix **L** such that $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathbf{T}}$, which means that **A** can be factored into $\mathbf{L}\mathbf{L}^{\mathbf{T}}$, where **L** is a lower triangular matrix with positive diagonal elements (Leon, 2006, p.389). Nonetheless, *R*-Package `Amelia II` does not allow estimating the ratio imputation model, nor do any of the existing multiple imputation software programs mentioned above.

Single Ratio Imputation

Suppose that the population model is equation (1). Under the following special case, the ratio $\bar{Y}_1 / \bar{Y}_2$ is an unbiased estimator of β, where ε_i is independent of Y_{i2} with the mean of 0 and the unknown variance of $Y_{i2}\sigma^2$ (Takahashi et al., 2017; Cochran, 1977; Shao, 2000; Liang et al., 2008). Under the general case, the ratio $\bar{Y}_1 / \bar{Y}_2$ is a consistent but biased estimator of β, and the mean of ε_i is 0 with unknown variance. However, as the sample size increases, this bias tends to be negligible. Also, the distribution of the ratio estimate is known to be asymptotically normal (Cochran, 1977, p.153).

$$Y_{i1} = \beta Y_{i2} + \varepsilon_i \tag{1}$$

Suppose Y_{it} is missing in the survey and that Y_{it-1} is fully observed in a previous dataset, where Y_{it} is the current value of the variable and Y_{it-1} is the value of the same variable at an earlier moment. The missing values of Y_t may be imputed by equation (2), where the value of β reflects the trend between the two time points.

$$\hat{Y}_{it} = \beta Y_{it-1} \tag{2}$$

A special case of equation (2) is cold deck imputation (de Waal et al., 2011), an example of which is that a missing value for unit *i* in an economic survey at *t* is

replaced with an observed value for unit i in another highly reliable dataset such as tax data at $t-1$. This model implies that the imputer is confident that β is always 1. Thus, there will be no estimation uncertainty whatsoever. A general case of equation (2) is ratio imputation (de Waal et al., 2011), an example of which is that a missing value for unit i of an economic survey at t is replaced with an observed value for unit i of the same economic survey at $t-1$, assuming that unit i answered at $t-1$. In this case, the imputer is not confident that β is always 1. Thus, there will be estimation uncertainty.

Therefore, in the general case of equation (2), the value of β is not known and must be estimated from the observed part of data. For this purpose, ratio imputation takes the form of a simple regression model without an intercept, whose slope coefficient is calculated not by OLS, but by the ratio between the means of the two variables. In other words, the ratio imputation model is equation (3), where $\hat{\beta} = \bar{Y}_{1,obs} / \bar{Y}_{2,obs}$. Also, ratio imputation can be made stochastic by adding a disturbance term as in equation (4) (Hu et al., 2001).

$$\hat{Y}_{i1} = \hat{\beta} Y_{i2} \tag{3}$$

$$\hat{Y}_{i1} = \hat{\beta} Y_{i2} + \hat{\varepsilon}_i \tag{4}$$

To illustrate, consider Table 1, where simulated data on income among 10 people are recorded. *Income0* is the unobserved truth, *Income1* is the current value, and *Income2* is the previous value. The mean of *Income0* is 504.500, the mean of *Income1* is 412.571, and the mean of *Income2* is 445.600.

Table 1. Example Data (Simulated Weekly Income in U.S. Dollars)

ID	*Income0*	*Income1*	*Income2*
1	543	543	514
2	272	272	243
3	**797**	**NA**	597
4	239	239	264
5	415	415	350
6	371	371	346
7	**650**	**NA**	545
8	495	495	475
9	553	553	564
10	**710**	**NA**	558

Note. *Income0* is the true complete variable. *Income1* is the observed incomplete variable with NA = missing. *Income2* is the auxiliary variable.

Presented in Table 2 are the imputed dataset by both deterministic ratio imputation and stochastic ratio imputation. The true model is, $\widehat{Income}_0 = \beta \times Income_2$ where $\beta = mean(Income_0) / mean(Income_2) = 1.132$. On the other hand, the imputation model is $\widehat{Income}_1 = \hat{\beta} \times Income_2$ where $\hat{\beta} = mean(Income_{1,obs}) / mean(Income_{2,obs}) = 1.048$. This clearly means that the imputation model consistently underestimates the true model due to missing values.

Table 2. Example of Imputed Data (Simulated Weekly Income in U.S. Dollars)

ID	*Income0*	*Income1*	Deterministic Ratio Imputation	Stochastic Ratio Imputation
1	543	543	543.000	543.000
2	272	272	272.000	272.000
3	**797**	**NA**	**625.594**	**586.441**
4	239	239	239.000	239.000
5	415	415	415.000	415.000
6	371	371	371.000	371.000
7	**650**	**NA**	**571.103**	**575.654**
8	495	495	495.000	495.000
9	553	553	553.000	553.000
10	**710**	**NA**	**584.756**	**621.730**

Note. *Income0* is the true complete variable. *Income1* is the observed incomplete variable with NA = missing.

The deterministic imputations are the exact predicted values by the imputation model. The stochastic imputations deviate from the predictions, reflecting fundamental uncertainty captured by $\hat{\varepsilon}_i$. Nevertheless, both types of ratio imputation models suffer from the lack of mechanism to incorporate estimation uncertainty, i.e., both models share the same deterministically calculated value of $\hat{\beta} = 1.048$, which is clearly different from the true $\beta = 1.132$.

Ratio imputation is considered to be an important tool in official statistics, because the model is supposed to be intuitively easy to verify for the practitioners (Bechtel et al., 2011). As a result, many national statistical agencies use ratio imputation in their statistical production processes, such as the U.S. Census Bureau (Thompson & Washington, 2012), the UK Office for National Statistics (2014), and Statistics Netherlands (de Waal et al., 2011), to name a few. However, this section demonstrated that the standard single ratio imputation models ignored estimation uncertainty. On this point, multiple ratio imputation comes to the rescue.

Theory of Multiple Ratio Imputation

If the missing mechanism is MAR, imputation can ameliorate the bias due to missingness (Little & Rubin, 2002; de Waal et al., 2011). Caution is required because imputed values are not the complete reproduction of the true values, and that the goal of imputation is generally not to replicate the truth for each missing value, but to make it possible to have a valid statistical inference. For this purpose, it is necessary to evaluate the error due to missingness, for which Rubin (1978, 1987) proposed multiple imputation as a solution. Indeed, Baraldi & Enders (2010) and Cheema (2014) demonstrated multiple imputation is superior to listwise deletion, mean imputation, and single regression imputation. Furthermore, Leite & Beretvas (2010) contended multiple imputation is robust to violations of continuous variables and the normality assumption. Thus, multiple imputation is the gold standard of treating missing data. The purpose of the current study, therefore, is to extend the utility of ratio imputation by transforming it to multiple imputation by way of the EMB algorithm described in this section.

Multiple imputation in theory is to randomly draw several imputed values from the distribution of missing data. However, missing data are by definition unobserved; as a result, the true distribution of missing data is always unknown. A solution to this problem is to estimate the posterior distribution of missing data based on observed data, and to make a random draw of imputed values. Honaker & King (2010) and Honaker et al. (2011) suggested the use of the EMB algorithm for the purpose of drawing the mean vector and the variance-covariance matrix from the posterior density, and presented a general-purpose multiple imputation software program called `Amelia II`, which is a computationally efficient and highly reliable multiple imputation program. Nevertheless, as presented above, `Amelia II` does not allow us to estimate the ratio imputation model.

The value of β was estimated by $\hat{\beta} = \bar{Y}_{1,obs} / \bar{Y}_{2,obs}$. Therefore, in order to create multiple ratio imputation, the mean vector needs to be randomly drawn from the posterior distribution of missing data given observed data. In the following sections, the EMB algorithm is applied to ratio imputation to create multiple ratio imputation. First, however, a review of the bootstrap method and the Expectation-Maximization (EM) algorithm is in order, to illustrate how the EMB algorithm works for the purpose of generating multiple ratio imputation.

Nonparametric Bootstrap

The first step for multiple ratio imputation is to randomly draw vectors of means from an appropriate posterior distribution to account for the estimation uncertainty. The EMB algorithm replaces the complex process of random draws from the posterior by nonparametric bootstrapping, which uses the existing sample data (size = n) as the pseudo-population and draws resamples (size = n) with replacement M times (Horowitz, 2001). If data $Y_1,\ldots,Y_n$ are independently and identically distributed from an unknown distribution F, this distribution is estimated by $\hat{F}(y)$, which is the empirical distribution F_n defined in equation (5), where $I(Y)$ is the indicator function of the set Y.

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y). \tag{5}$$

Based on equation (5), bootstrap resamples are generated. The distribution $\hat{F}$ can be any estimator in order to generate the bootstrap resamples of F based on $Y_1,\ldots,Y_n$. A nonparametric estimator of F is the empirical distribution F_n defined by equation (5) (Shao & Tu, 1995, pp. 2-4, pp. 9-11; DeGroot & Schervish, 2002, pp.753-754).

Table 3. Bootstrap Data (M = 2)

Incomplete Data		Bootstrap 1		Bootstrap 2	
Income1	***Income2***	***IncomeB11***	***IncomeB12***	***IncomeB21***	***IncomeB22***
543	514	NA	545	495	475
272	243	272	243	272	243
NA	597	239	264	371	346
239	264	NA	597	415	350
415	350	272	243	NA	597
371	346	553	564	543	514
NA	545	272	243	272	243
495	475	495	475	NA	545
553	564	553	564	371	346
NA	558	272	243	NA	545

Note. NA represents missing values.

This is illustrated in Table 3. The incomplete data are the original missing data in Table 1. When listwise deletion is applied to this dataset, the mean of

Income1 is 412.571. The Bootstrap 1 and Bootstrap 2 in Table 3 refer to the bootstrap resamples, where $M = 2$. When listwise deletion is applied to these bootstrap datasets, the mean of *IncomeB11* is 366.000 and the mean of *IncomeB21* is 391.286. The variation between these estimates is the essential mechanism of capturing estimation uncertainty due to imputation.

However, when incomplete data are bootstrapped, the chance is that each bootstrap resample is also incomplete. Therefore, the information from incomplete bootstrap resamples is biased and inefficient. The EM algorithm refines bootstrap estimates in the next section.

EM Algorithm

MLEs are the parameter estimates that maximize the likelihood of observing the existing data (Long, 1997, p.26), which have the NICE properties of asymptotic Normality, Invariance, Consistency, and asymptotic Efficiency (Greene, 2003). Nevertheless, it is difficult to directly calculate MLE in missing data. Making incomplete data complete requires information about the distribution of the data, such as the mean and the variance-covariance; however, these incomplete data are used to estimate the mean and the variance-covariance. Therefore, it is not straightforward to analytically solve this problem. For the purpose of dealing with this problem, iterative methods such as the EM algorithm were proposed to estimate such quantities of interest (Allison, 2002).

A certain distribution is assumed in the EM algorithm, as are tentative starting values for the mean and the variance-covariance. An expected value of model likelihood is calculated, the likelihood is maximized, model parameters are estimated that maximize these expected values, and then the distribution is updated. The expectation and the maximization steps are repeated until the values converge, whose properties are known to be an MLE (Schafer, 1997; Iwasaki, 2002; Do & Batzoglou, 2008). Formally, the EM algorithm can be summarized as follows. Starting from an initial value θ_0, repeat the following two steps:

1. E-step: $Q(\theta|\theta_t) = \int l(\theta|Y) P(Y_{mis}|Y_{obs};\theta_t)\, dY_{mis}$, where $l(\theta|Y)$ is log likelihood.
2. M-step: Maximize $\theta_{t+1} = \arg\max_\theta Q(\theta|\theta_t)$ with respect to θ.

Under certain conditions, it is proven that $\theta_t \to \hat{\theta}(t \to \infty)$.

The values in Table 3 were incomplete. If the EM algorithm is used to refine these values, the EM mean for *IncomeB11* is 405.741 and the EM mean for *IncomeB12* is 398.100; also, the EM mean for *IncomeB21* is 450.912 and the EM mean for *IncomeB22* is 420.400. Using these values, the ratio will be estimated as 1.019 and 1.072, respectively. Thus, in this small example, the ratio is estimated as 1.046 on average, ranging from 1.019 to 1.072. This variation captures the estimation uncertainty due to missingness, which is called the between-imputation variance (Little & Rubin, 2002). Obviously, real applications require a much larger value of *M* (Graham et al., 2007; Bodner, 2008).

Application of the EMB Algorithm to Multiple Ratio Imputation

The multiple ratio imputation model is defined by equation (6), where tilde means that these values are drawn from an appropriate posterior distribution of missing data. In other words, $\tilde{\beta}$ is a vector of ratios drawn from the appropriate posterior taking estimation uncertainty into account and $\tilde{\varepsilon}_i$ is the disturbance term taking fundamental uncertainty into account (King et al., 2001).

$$\tilde{Y}_{i1} = \tilde{\beta} Y_{i2} + \tilde{\varepsilon}_i, \text{ where } \tilde{\beta} = \frac{\tilde{\bar{Y}}_1}{\tilde{\bar{Y}}_2} \tag{6}$$

Table 4. Multiple Ratio Imputation Data (M = 2)

ID	*Income1*	*Income2*	*Imputation1*	*Imputation2*
1	543	514	543.000	543.000
2	272	243	272.000	272.000
3	**NA**	597	**620.917**	**662.732**
4	239	264	239.000	239.000
5	415	350	415.000	415.000
6	371	346	371.000	371.000
7	**NA**	545	**571.100**	**600.655**
8	495	475	495.000	495.000
9	553	564	553.000	553.000
10	**NA**	558	**597.406**	**637.115**

Presented in Table 4 are the result of multiple ratio imputation, where $M = 2$, using the same example data as in Table 1. The model is $Income_1 = \tilde{\beta} \times Income_2 + \tilde{\varepsilon}_i$. If $M = 100$, the mean of $\tilde{\beta}$ is 1.050 with a standard deviation of 0.048, ranging from 0.903 to 1.342. This variation captures the

stability of the imputation model, which serves as a diagnostic method for imputation, because the simulation standard error (between-imputation variance) can be appropriately used for assessing the likeliness of the simulation estimator being close to the true parameter of interest (DeGroot & Schervish, 2002). In Table 4, the values of *Imputation1* and *Imputation2* for ID 3, 7, and 10 change over columns *Imputation1* to *Imputation2*, because the values in these rows are imputed values. Also, note that the values in the other rows do not change over columns, because they are observed values.

Just as in regular multiple imputation (Little & Rubin, 2002), the estimates by multiple ratio imputation can be combined as follows. Let $\hat{\theta}_m$ be an estimate based on the m^{th} multiple-imputed dataset. The combined point estimate $\bar{\theta}_M$ is equation (7).

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{7}$$

The variance of the combined point estimate consists of two parts. Let v_m be the estimate of the variance of $\hat{\theta}_m$, var($\hat{\theta}_m$), let $\bar{W}_M$ be the average of within-imputation variance, let $\bar{B}_M$ be the average of between-imputation variance, and let T_M be the total variance of $\bar{\theta}_M$. Then, the total variance of $\bar{\theta}_M$ is equation (8), where $(1 + 1/M)$ is an adjustment factor because M is not infinite. If M is infinite, $\lim_{M\to\infty}\left(1+\frac{1}{M}\right)\tilde{v}_M = \tilde{v}_M$. In short, the variance of $\bar{\theta}_M$ takes into account within-imputation variance and between-imputation variance.

$$T_M = \bar{W}_M + \left(1+\frac{1}{M}\right)\bar{B}_M = \frac{1}{M}\sum_{m=1}^{M} v_m + \left(1+\frac{1}{M}\right)\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \bar{\theta}_M\right)^2\right] \tag{8}$$

Graphically outlined in Figure 1 is a schematic overview of multiple ratio imputation ($M = 5$). In summary, multiple ratio imputation replaces missing values by M simulated values, where $M > 1$. Conditional on observed data, the imputer constructs a posterior distribution of missing data, draws a random sample from this distribution, and creates several imputed datasets. Then, conduct the standard statistical analysis, separately using each of the M multiple-imputed datasets, and combine the results of the M statistical analyses in the above manner to calculate a point estimate just as in regular multiple imputation.

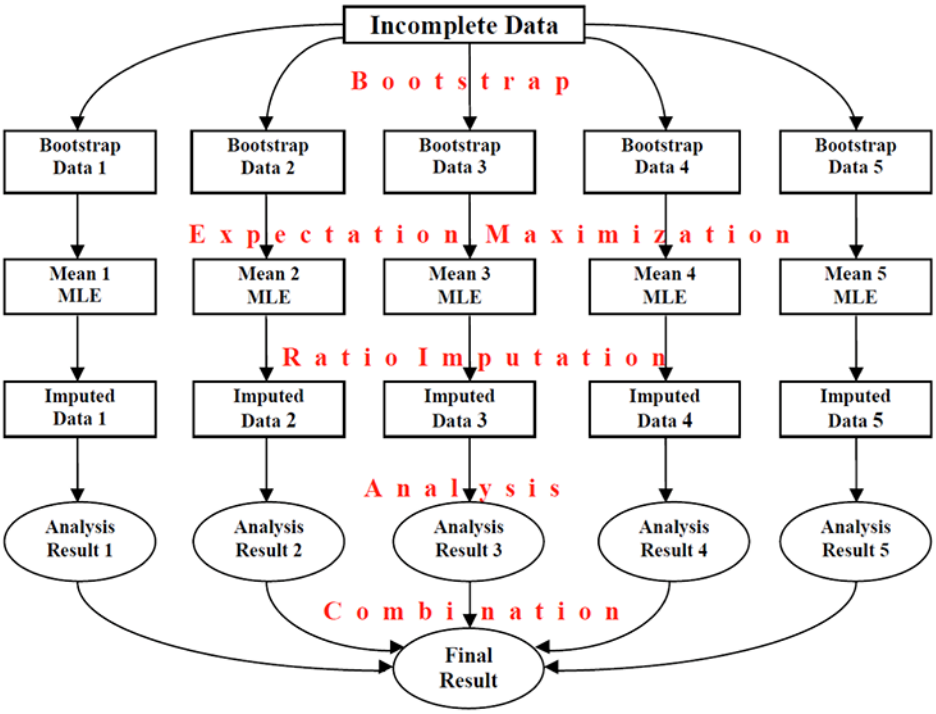


Figure 1. Schematic of Multiple Ratio Imputation by the EMB Algorithm ($M = 5$)

Monte Carlo Evidence

Using 45,000 simulated datasets with various characteristics, the Relative Root Mean Square Errors (RRMSE) of the estimators for the mean, the standard deviation, and the *t*-statistics in regression across different missing data handling techniques are compared. The data are a modified version of the simulated data used by King et al. (2001). The Monte Carlo experiments are based on 1,000 iterations, each of which is a random draw from the following multivariate normal distribution: Variables $y1$ and $y2$ are normally distributed with the mean vector (6, 10) and the standard deviation vector (1, 1), where the correlation between $y1$ and $y2$ is set to 0.6 (Note that the value of 0.6 was chosen because this is approximately the correlation value among the variables in official economic statistics which is the target of the current study. Also, in other few runs, not reported, the parameter values were changed, and the conclusions were very similar). Each set of these 1,000 data is repeated for $n = 50$, $n = 100$, $n = 200$, $n = 500$, and $n = 1{,}000$; thus, there are 5,000 datasets of five different data sizes. Our simulated data assume that the population model is equation (9).

$$Y_{i1} = \beta Y_{i2} + \varepsilon_i, \text{ where } \beta = \frac{\bar{Y}_1}{\bar{Y}_2} = 0.6, \varepsilon_i \sim N\left(0, 0.64\right). \quad (9)$$

Furthermore, following King et al. (2001), each of these 5,000 datasets is made incomplete using the three data generation processes of MCAR, MAR, and NI as in Table 5. Under the assumption of MCAR, the missingness of *y*1 randomly depends on the values of u (uniform random numbers). Under the assumption of MAR, the missingness of *y*1 depends on the values of *y*2 and *u*. Under the assumption of NI, the missingness of *y*1 depends on the observed and unobserved values of *y*1 itself and the values of *u*.

Table 5. Missingness Mechanisms and Missing Rates

	Missingness of *y*1 is a function of *u*.
MCAR	15%: *y*1 is missing if *u* > 0.85.
	25%: *y*1 is missing if *u* > 0.75.
	35%: *y*1 is missing if *u* > 0.65.
	Missingness of *y*1 is a function of *y*2 and *u*.
MAR	15%: *y*1 is missing if *y*2 > 10 and *u* > 0.7.
	25%: *y*1 is missing if *y*2 > 10 and *u* > 0.5.
	35%: *y*1 is missing if *y*2 > 10 and *u* > 0.3.
	Missingness of *y*1 is a function of *y*1, *x*, and *u*.
NI	15%: *y*1 is missing if *y*1 > 6 and *u* > 0.7.
	25%: *y*1 is missing if *y*1 > 6 and *u* > 0.5.
	35%: *y*1 is missing if *y*1 > 6 and *u* > 0.3.

Variable *y*1 is the target incomplete variable for imputation, Variable *y*2 is completely observed in all of the situations to be used as the auxiliary variable, and Variable *u* in Table 5 is 1,000 sets of continuous uniform random numbers ranging from 0 to 1 for the missingness mechanism. The average missing rates are set to 15%, 25%, and 35%. These missing rates approximately cover the range from 10% to 40% missingness.

The performance can be captured by the Mean Square Error (MSE), defined as equation (10), where θ is the true quantity of interest and $\hat{\theta}$ is an estimator. The MSE measures the dispersion around the true value of the parameter, suggesting that an estimator with the smallest MSE is the best of a competing set of estimators (Gujarati, 2003, p. 901).

$$MSE\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right)^2 \tag{10}$$

For the ease of interpretation, following Di Zio & Guarnera (2013), the Relative Root Mean Square Error (RRMSE) is used, which is defined as equation (11), where θ is the truth, $\hat{\theta}$ is an estimator, and T is the number of trials. For example, θ in the following analyses is the mean, the standard deviation, and the *t*-statistic based on complete data. $\hat{\theta}$ is the estimated quantity based on imputed data. T is 1,000.

$$RRMSE\left(\hat{\theta}\right) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{\theta} - \theta}{\theta}\right)^2} \tag{11}$$

The complete results based on the 45,000 datasets are presented in Tables 6, 8, and 9. In the following analyses, the multiple ratio imputation model sets the number of multiple-imputed datasets (M) to 100, based on the recent findings in the multiple imputation literature (Graham et al., 2007; Bodner, 2008).

RRMSE Comparisons for the Mean

Presented in Table 6 are the RRMSE comparisons for the mean among listwise deletion, deterministic single ratio imputation, and multiple ratio imputation (M = 100), where the RRMSE is averaged over the 1,000 simulations. For multiple ratio imputation, the 100 mean values are combined using equation (7) in each of the 1,000 simulations.

The standard recommendation (de Waal et al., 2011, p.245) is that if the goal is to calculate a point estimate, the choice is deterministic single ratio imputation. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is as good as that of deterministic single ratio imputation, which is known to be a preferred method for the estimation of the mean. If multiple ratio imputation equally performs well compared to deterministic single ratio imputation, this means that multiple ratio imputation attains the highest performance in estimating the mean.

Table 6. RRMSE Comparisons for the Mean (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Deterministic Ratio Imputation	Multiple Ratio Imputation
50	15%	*MCAR*	0.009	0.008	0.008
		MAR	0.017	0.008	0.008
		NI	0.026	0.017	0.018
	25%	*MCAR*	0.014	0.011	0.011
		MAR	0.03	0.01	0.011
		NI	0.048	0.032	0.033
	35%	*MCAR*	0.017	0.014	0.014
		MAR	0.045	0.012	0.014
		NI	0.075	0.05	0.052
100	15%	*MCAR*	0.007	0.006	0.006
		MAR	0.016	0.005	0.005
		NI	0.024	0.016	0.016
	25%	*MCAR*	0.01	0.008	0.008
		MAR	0.028	0.007	0.008
		NI	0.046	0.03	0.03
	35%	*MCAR*	0.012	0.01	0.01
		MAR	0.044	0.008	0.01
		NI	0.073	0.048	0.05
200	15%	*MCAR*	0.005	0.004	0.004
		MAR	0.015	0.004	0.004
		NI	0.024	0.016	0.016
	25%	*MCAR*	0.007	0.005	0.005
		MAR	0.028	0.005	0.005
		NI	0.045	0.029	0.03
	35%	*MCAR*	0.009	0.007	0.007
		MAR	0.043	0.006	0.007
		NI	0.072	0.048	0.049
500	15%	*MCAR*	0.003	0.003	0.003
		MAR	0.014	0.002	0.002
		NI	0.024	0.015	0.015
	25%	*MCAR*	0.004	0.003	0.003
		MAR	0.027	0.003	0.003
		NI	0.045	0.029	0.029
	35%	*MCAR*	0.006	0.004	0.004
		MAR	0.043	0.004	0.005
		NI	0.072	0.047	0.048
1000	15%	*MCAR*	0.002	0.002	0.002
		MAR	0.014	0.002	0.002
		NI	0.024	0.015	0.015
	25%	*MCAR*	0.003	0.003	0.003
		MAR	0.027	0.002	0.002
		NI	0.044	0.029	0.029
	35%	*MCAR*	0.004	0.003	0.003
		MAR	0.043	0.002	0.003
		NI	0.072	0.047	0.048

Note. Average over the 1,000 simulations for each data type. M = 100 for multiple ratio imputation

In 42 of the 45 patterns, deterministic ratio imputation and multiple imputation both outperform listwise deletion with 3 ties. Even when the missing mechanism is MCAR, the results by imputation are almost always better than those of listwise deletion. Between the ratio imputation methods, deterministic ratio imputation slightly performs better than multiple ratio imputation in 14 out of the 45 patterns with 31 ties. However, the largest difference is only 0.002 in terms of the RRMSE. Thus, there are no significant differences between deterministic ratio imputation and multiple ratio imputation. Furthermore, this difference is expected to completely disappear as M approaches infinity. In general, under the situations where the model is correctly specified and the assumption of MAR is satisfied, both single imputation and multiple imputation ($M = \infty$) would be unbiased and agree on the point estimation (Donders et al., 2006). The results in Table 6 ensure this general relationship also applies to the relationship between single ratio imputation and multiple ratio imputation. Therefore, on average, multiple ratio imputation can be expected to give essentially the same answers as to the estimation of the mean, compared to deterministic ratio imputation.

Multiple ratio imputation can be more useful than deterministic single ratio imputation in the estimation of the mean, because multiple ratio imputation has more information in its output. Recall that there are three sources of variation in multiple imputation (van Buuren, 2012). One is the conventional measure of statistical variability (also known as within-imputation variance). Another is the additional variance due to missing values in the data (also known as between-imputation variance). The last one is simulation variance by the finite number of multiple-imputed data captured by $\bar{B}_M / M$ in equation (8). Among these, the between-imputation variance is particularly important, because it reflects the uncertainty associated with missingness (Honaker et al., 2011).

To demonstrate how multiple ratio imputation provides additional information on the between-imputation variance, presented in Table 7 is the mean of $y1$ when the missing data mechanism is MAR with the average missing rate of 35%, where the reported values are the average over the 1,000 simulations. In Table 7, when the missing data mechanism is MAR, both of the imputation methods are almost equally accurate, in terms of estimating the mean. Additionally, multiple ratio imputation has more rows in Table 7 for BISD and CI (95%). BISD stands for the Between-Imputation Standard Deviation, and CI (95%) stands for the Confidence Interval associated with estimation error due to missingness at the 95% level. BISD is the square-root of the between-imputation

variance and measures the dispersion of the 100 mean values based on multiple ratio imputation ($M = 100$). In other words, BISD is the variation in the distribution of the estimated mean, which is usually called the standard error (Baraldi & Enders, 2010, p.16). Thus, based on BISD, the imputer can be approximately 95% confident that the true mean value of complete data is somewhere between 5.941 and 6.057, after taking the error due to missingness into account. Furthermore, the imputer can be approximately 95% confident that the imputed mean value (6.00) is meaningfully different from the listwise deletion estimate (5.74), which is outside the 95% confidence interval (5.94, 6.06). Single ratio imputation (both deterministic and stochastic) lacks this mechanism of assessing estimation uncertainty.

Table 7. Mean of *y*1 (MAR-35%)

	Complete Data	Listwise Deletion	Deterministic Ratio Imputation	Multiple Ratio Imputation
Mean	6.000	5.741	6.000	5.999
BISD	NA	NA	NA	0.029
CI (95%)	NA	NA	NA	5.941, 6.057
n	500	325	500	500

Note. NA means Not-Applicable. Average over the 1,000 simulations. M = 100 for multiple ratio imputation

RRMSE Comparisons for the Standard Deviation

Presented in Table 8 are the RRMSE comparisons for the standard deviation among listwise deletion, stochastic single ratio imputation, and multiple ratio imputation ($M = 100$), where the RRMSE is averaged over the 1,000 simulations. For multiple ratio imputation, the 100 standard deviation values are combined using equation (7) in each of the 1,000 simulations.

The standard recommendation (de Waal et al., 2011) is that if the goal is to estimate the variation of data, the choice is stochastic single ratio imputation. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is as good as that of stochastic ratio imputation, which is known to be a preferred method to estimate the standard deviation. Note that, in other simulation runs, the EM algorithm was applied to the imputed data by the deterministic ratio imputation model, in order to compute the standard deviation. However, these results were not good and thus omitted here.

Table 8. RRMSE Comparisons for the Standard Deviation (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Stochastic Ratio Imputation	Multiple Ratio Imputation
50	15%	*MCAR*	0.042	0.048	0.037
		MAR	0.045	0.047	0.038
		NI	0.048	0.052	0.043
	25%	*MCAR*	0.059	0.062	0.049
		MAR	0.066	0.062	0.054
		NI	0.079	0.074	0.067
	35%	*MCAR*	0.075	0.075	0.058
		MAR	0.088	0.071	0.067
		NI	0.146	0.117	0.118
100	15%	*MCAR*	0.029	0.035	0.026
		MAR	0.031	0.034	0.026
		NI	0.035	0.037	0.031
	25%	*MCAR*	0.040	0.044	0.033
		MAR	0.046	0.044	0.037
		NI	0.064	0.058	0.054
	35%	*MCAR*	0.052	0.052	0.040
		MAR	0.067	0.054	0.047
		NI	0.121	0.097	0.098
200	15%	*MCAR*	0.021	0.025	0.018
		MAR	0.022	0.025	0.019
		NI	0.025	0.027	0.023
	25%	*MCAR*	0.028	0.030	0.023
		MAR	0.036	0.032	0.027
		NI	0.049	0.044	0.042
	35%	*MCAR*	0.037	0.037	0.028
		MAR	0.053	0.038	0.034
		NI	0.109	0.086	0.088
500	15%	*MCAR*	0.014	0.016	0.012
		MAR	0.014	0.016	0.012
		NI	0.018	0.019	0.016
	25%	*MCAR*	0.018	0.020	0.015
		MAR	0.024	0.020	0.017
		NI	0.042	0.038	0.036
	35%	*MCAR*	0.022	0.023	0.018
		MAR	0.043	0.024	0.021
		NI	0.106	0.083	0.084
1000	15%	*MCAR*	0.010	0.012	0.008
		MAR	0.010	0.011	0.008
		NI	0.014	0.015	0.013
	25%	*MCAR*	0.013	0.014	0.011
		MAR	0.019	0.014	0.011
		NI	0.040	0.037	0.033
	35%	*MCAR*	0.017	0.017	0.013
		MAR	0.038	0.016	0.014
		NI	0.100	0.080	0.079

Note. Average over the 1,000 simulations for each data type. M = 100 for multiple ratio imputation

In all of the 45 patterns, multiple ratio imputation always outperforms listwise deletion. Even when the missing mechanism is MCAR, the results by multiple ratio imputation are always better than those of listwise deletion. In contrast, stochastic ratio imputation outperforms listwise deletion in only 20 out of the 45 patterns. Especially, when the missing mechanism is MCAR, listwise deletion often outperforms stochastic ratio imputation in 11 out of the 15 patterns with 4 ties, although the difference is minimal. This implies that when missing data are suspected to be MCAR, there is a chance that using stochastic ratio imputation may make the situation worse than simply using listwise deletion. When the missing mechanism is MAR or NI, stochastic ratio imputation indeed outperforms listwise deletion in 20 out of the 30 patterns.

Between the ratio imputation methods, multiple ratio imputation often performs better than stochastic ratio imputation, 41 out of the 45 patterns. Therefore, this study contends that multiple ratio imputation is the preferred method for the estimation of the standard deviation. Table 8 implies that, regardless of missing mechanisms, multiple ratio imputation should be used for the purpose of estimating the standard deviation.

Just as in the case of estimating the mean, let us take the case of 35% missingness with the MAR condition as an example. Based on BISD, the imputer can be approximately 95% confident that the true standard deviation value of complete data is somewhere between 0.960 and 1.040, after taking the error due to missingness into account.

RRMSE Comparisons for the *t*-Statistics in Regression

The comparisons in this section are particularly important because even if the intercept should be zero and the slope should be estimated by the ratio between two variables, there are no other choices but to stick to regular multiple imputation for the computation of the *t*-statistics in regression. The regression model in Table 9 is $y2 = a + b*y1$. The quantity of interest is the *t*-statistic of *b*, i.e., $t_b = b/se(b)$. The RRMSE reported here measures the average distance between the true t_b based on complete data and the estimated t_b based on imputed data. Table 9 presents the RRMSE comparisons for the *t*-statistics in regression among listwise deletion, regular multiple imputation (`Amelia II`), and multiple ratio imputation, where $M = 100$ for both regular multiple imputation and multiple ratio imputation, and the RRMSE is averaged over the 1,000 simulations. For regular multiple imputation and multiple ratio imputation, the 100 coefficient values are combined using equation (7), the 100 standard error values are

combined using equation (8), and the *t*-statistics are calculated using these two values in each of the 1,000 simulations.

Remember that the multiple ratio imputation model is equation (6). On the other hand, multiple imputation by `Amelia II` is equation (12), where the coefficients are random draws of the mean vectors and the variance-covariance matrices from the posterior distribution (Honaker & King, 2010).

$$\tilde{Y}_{i1} = \tilde{\beta}_0 + \tilde{\beta}_1 Y_{i2} + \tilde{\varepsilon}_i \text{, where } \tilde{\beta}_1 = \frac{\text{cov}\left(Y_{i1}, Y_{i2}\right)}{\text{var}\left(Y_{i2}\right)}, \tilde{\beta}_0 = \tilde{\bar{Y}}_1 - \tilde{\beta}_1 \tilde{\bar{Y}}_2 . \tag{12}$$

The standard recommendation (van Buuren, 2012; Hughes et al., 2014) is that if the goal is to obtain valid inferences with standard errors, the choice is multiple imputation which is a superior variance-estimation method. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is better than that of regular multiple imputation in terms of estimating the *t*-statistics. The comparison of the *t*-statistics in regression is appropriate, because it is the quantity of interest for many applied researchers in disputing whether an independent variable has some impact on a dependent variable. According to Cheema (2014), comparisons of *t*-statistics are fair because the complete sample and the imputed sample are identical in all respects including power, except for the fact that no values were missing in the complete sample while some values were missing in the imputed values. Therefore, the differences in the observed values of statistics are caused by the differences between imputed values and their true counterparts.

The comparison of multiple ratio imputation and `Amelia II` is appropriate, because the algorithm is the same EMB under the same platform of the *R* statistical environment. In all of the 45 patterns, regular multiple imputation and multiple ratio imputation both outperform listwise deletion. Furthermore, multiple ratio imputation almost always outperforms regular multiple imputation 43 out of the 45 patterns under the condition where the true population model is equation (9). Thus, when the true model is a ratio model such as equation (9), multiple ratio imputation is more accurate and efficient than regular multiple imputation.

Therefore, multiple ratio imputation adds an important option for the tool kit of imputing and analyzing the mean, the standard deviation, and the *t*-statistics. If the true model is equation (9), multiple ratio imputation is at least as good as and in many cases better than the other traditional imputation methods for the three quantities of interest, regardless of the missingness mechanisms. However, it is

Table 9. RRMSE Comparisons for *t*-statistics (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Multiple Imputation Amelia II	Multiple Ratio Imputation
50	15%	*MCAR*	0.126	0.103	0.087
		MAR	0.137	0.107	0.093
		NI	0.141	0.114	0.099
	25%	*MCAR*	0.185	0.144	0.113
		MAR	0.220	0.173	0.135
		NI	0.222	0.175	0.138
	35%	*MCAR*	0.242	0.189	0.134
		MAR	0.317	0.247	0.171
		NI	0.328	0.269	0.179
100	15%	*MCAR*	0.104	0.075	0.066
		MAR	0.113	0.080	0.071
		NI	0.111	0.081	0.072
	25%	*MCAR*	0.159	0.109	0.087
		MAR	0.192	0.127	0.101
		NI	0.194	0.136	0.108
	35%	*MCAR*	0.218	0.153	0.107
		MAR	0.294	0.191	0.131
		NI	0.297	0.224	0.147
200	15%	*MCAR*	0.091	0.059	0.052
		MAR	0.101	0.064	0.056
		NI	0.101	0.066	0.060
	25%	*MCAR*	0.145	0.092	0.075
		MAR	0.181	0.106	0.085
		NI	0.177	0.117	0.095
	35%	*MCAR*	0.208	0.136	0.097
		MAR	0.282	0.159	0.113
		NI	0.282	0.199	0.133
500	15%	*MCAR*	0.084	0.050	0.044
		MAR	0.094	0.053	0.047
		NI	0.093	0.058	0.051
	25%	*MCAR*	0.141	0.086	0.066
		MAR	0.171	0.092	0.069
		NI	0.170	0.107	0.083
	35%	*MCAR*	0.202	0.127	0.086
		MAR	0.279	0.144	0.097
		NI	0.282	0.193	0.121
1000	15%	*MCAR*	0.080	0.046	0.041
		MAR	0.089	0.046	0.043
		NI	0.091	0.048	0.049
	25%	*MCAR*	0.137	0.053	0.063
		MAR	0.167	0.084	0.067
		NI	0.168	0.105	0.083
	35%	*MCAR*	0.198	0.122	0.084
		MAR	0.275	0.132	0.092
		NI	0.275	0.186	0.120

Note. Average over the 1,000 simulations for each data type. M = 100 for multiple imputation

not claimed multiple ratio imputation is always superior to regular multiple imputation. If the true model is not a ratio model such as equation (9), the superiority shown in this section is not guaranteed.

Conclusion

A novel application of the EMB algorithm to ratio imputation was proposed, along with the mechanism and the usefulness of multiple ratio imputation. Monte Carlo evidence was presented, where the newly-developed *R*-function called `MrImputation` (Takahashi, 2017) for multiple ratio imputation was applied to the 45,000 simulated data.

It was shown the fit of multiple ratio imputation was generally as good as or sometimes better than that of single ratio imputation and regular multiple imputation if the assumption holds. Specifically, for the purpose of estimating the mean, the performance of deterministic ratio imputation and multiple ratio imputation are essentially equally good, with multiple ratio imputation having additional information on estimation uncertainty. For the purpose of estimating the standard deviation, multiple ratio imputation outperforms stochastic ratio imputation. For the purpose of estimating the *t*-statistics in regression, multiple ratio imputation clearly outperforms regular multiple imputation when the population model is equation (9).

These findings are important because it is often recommended to use different ways of imputation depending on the type of statistical analyses, meaning that there are no one-size-fit-for-all imputation methods (Poston & Conde, 2014). Thus, multiple ratio imputation will be a valuable addition for treating missing data problems, so that multiple ratio imputation will expand the choice of missing data treatments.

This is only a starting point for multiple ratio imputation. There are three multiple imputation algorithms. The version of multiple ratio imputation introduced here used the Expectation-Maximization with Bootstrapping algorithm. However, multiple ratio imputation is a generic imputation model; thus, future research may apply the other two multiple imputation algorithms to expand the scope and the applicability of the method.

Acknowledgments

The author wishes to thank Dr. Manabu Iwasaki (Seikei University), Dr. Michiko Watanabe (Keio University), Dr. Takayuki Abe (Keio University), Dr. Tetsuto

Himeno (Shiaga University), and Mr. Nobuyuki Sakashita (Statistical Research and Training Institute) for their valuable comments on earlier versions of this article. The author also wishes to thank the two anonymous reviewers for useful comments to revise this article. However, any remaining errors are the author's responsibility. Also, note that the views and opinions expressed in this article are the author's own, not necessarily those of the institution. The analyses in this article were conducted using *R* 3.1.0.

References

Allison, P. D. (2002). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. doi: 10.1016/j.jsp.2009.10.001

Bechtel, L., Gonzalez, Y., Nelson, M., & Gibson, R. (2011). Assessing several hot deck imputation methods using simulated data from several economic programs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 5022-5036.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15*(4), 651-675. doi: 10.1080/10705510802339072

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple Imputation and its Application*. Chichester, West Sussex: John Wiley & Sons. doi: 10.1002/9781119942283

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods, 13*(2), 53-75.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed). New York, NY: John Wiley & Sons.

DeGroot, M H., & Schervish, M. J. (2002). *Probability and Statistics* (3rd ed). Boston, MA: Addison-Wesley.

de Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470904848

Di Zio, M., & Guarnera, U. (2013). Contamination model for selective editing. *Journal of Official Statistics, 29*(4), 539-555. doi: 10.2478/jos-2013-0039

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology, 26*(8), 897-899. doi: 10.1038/nbt1406

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087-1091. doi: 10.1016/j.jclinepi.2006.01.014

Fox, J. (2015). Package 'Norm' [Computer software]. Retrieved from: http://cran.r-project.org/web/packages/norm/norm.pdf

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213. doi: 10.1007/s11121-007-0070-9

Greene, W. A. (2003). *Econometric Analysis* (5th ed). Upper Saddle River, NJ: Prentice Hall.

Gujarati, D. N. (2003). *Basic econometrics* (4th ed). New York, NY: McGraw-Hill.

Honaker, J., & King, G. (2010). What to do about missing values in time series cross-section data. *American Journal of Political Science, 54*(2), 561-581. doi: 10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: a program for missing data. *Journal of Statistical Software, 45*(7), 1-47. doi: 10.18637/jss.v045.i07

Honaker, J., King, G., & Blackwell, M. (2015). Package 'Amelia' [Computer software]. Retrieved from: http://cran.r-project.org/web/packages/Amelia/Amelia.pdf

Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman & E. Leamer (Eds), *Handbook of Econometrics* (pp. 3160-3228), Vol. 5. Amsterdam: Elsevier. doi: 10.1016/s1573-4412(01)05005-x

Hu, M., Salvucci, S., & Lee, R. (2001). *A Study of Imputation Algorithms. Working Paper No. 2001–17*. U.S. Department of Education. National Center for Education Statistics. Retrieved from: http://nces.ed.gov/pubs2001/200117.pdf

Hughes, R. A., Sterne, J. A. C., & Tilling, K. (2014). Comparison of imputation variance estimators. *Statistical Methods in Medical Research*, *25*(6), 2541-2557. doi: 10.1177/0962280214526216

Iwasaki, M. (2002). *Fukanzen Data no Toukei Kaiseki (Foundations of Incomplete Data Analysis)*. Tokyo: EconomistSha Publications, Inc.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review, 95*(1), 49-69.

Leite, W., & Beretvas, S. (2010). The performance of multiple imputation for Likert-type items with missing data. *Journal of Modern Applied Statistical Methods, 9*(1), 64-74.

Leon, S. J. (2006). *Linear Algebra with Applications* (7th ed). Upper Saddle River, NJ: Pearson/Prentice Hall.

Liang, H., Su, H., & Zou, G. (2008). Confidence intervals for a common mean with missing data with applications in AIDS study. *Computational Statistics & Data Analysis, 53*(2), 546-553. doi: 10.1016/j.csda.2008.09.021

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781119013563

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Office for National Statistics. (2014). *Change to imputation method used for the turnover question in monthly business surveys. Guidance and methodology: retail sales.* Retrieved from: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/retail-sales/index.html

Poston, D., & Conde, E. (2014). Missing data and the statistical modeling of adolescent pregnancy. *Journal of Modern Applied Statistical Methods, 13*(2), 464-478.

Rubin, D. B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.

SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. Retrieved from: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.

Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by "Missing at Random"? *Statistical Science, 28*(2), 257-268.

Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology, 26*(1), 79-85.

Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap*. New York, NY: Springer. doi: 10.1007/978-1-4612-0795-5

SPSS Inc. (2009). *PASW Missing Values 18*. Retrieved from: http://www.unt.edu/rss/class/Jon/SPSS_SC/Manuals/v18/PASW Missing Values 18.pdf

Statistical Solutions. (2011). *SOLAS Version 4.0 Imputation User Manual*. Retrieved from: http://www.solasmissingdata.com/wp-content/uploads/2011/05/Solas-4-Manual.pdf

Takahashi, M. (2017). Implementing multiple ratio imputation by the EMB algorithm in *R. Journal of Modern Applied Statistical Methods, 16*(1). doi: 10.22237/jmasm/1493598900

Takahashi, M., & Ito, T. (2013). Multiple imputation of missing values in economic surveys: comparison of competing algorithms. *Proceedings of The 59th World Statistics Congress of the International Statistical Institute (ISI)*, 3240-3245.

Takahashi, M., Iwasaki, M., & Tsubaki, H. (2017). Imputing the mean of a heteroskedastic log-normal missing variable: A unified approach to ratio imputation. *Statistical Journal of the IAOS* (forthcoming). doi: 10.3233/sji-160306

Thompson, K. J., & Washington, K. T. (2012). A response propensity based evaluation of the treatment of unit nonresponse for selected business surveys. *Federal Committee on Statistical Methodology 2012 Research Conference*. Retrieved from: https://fcsm.sites.usa.gov/files/2014/05/Thompson_2012FCSM_III-B.pdf

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/b11826

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in *R. Journal of Statistical Software, 45*(3), 1-67. doi: 10.18637/jss.v045.i03

van Buuren, S., & Groothuis-Oudshoorn, K. (2015). Package 'mice' [Computer software]. Retrieved from: http://cran.r-project.org/web/packages/mice/mice.pdf

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 657-673.
doi: 10.22237/jmasm/1493598900

ISSN 1538 − 9472

JMASM44: Implementing Multiple Ratio Imputation by the EMB Algorithm (R)

Masayoshi Takahashi
Tokyo University of Foreign Studies
Tokyo, Japan

Although single ratio imputation is often used to deal with missing values in practice, there is a paucity of discussion regarding multiple ratio imputation. Code in the *R* statistical environment is presented to execute multiple ratio imputation by the Expectation-Maximization with Bootstrapping (EMB) algorithm.

Keywords: Multiple imputation, ratio imputation, Expectation-Maximization, bootstrap, missing data, incomplete data, nonresponse, estimation uncertainty

Introduction

Code is presented for multiple ratio imputation step by step in the imputation stage, followed by the analysis stage. The Appendix combines these *R*-codes to present Software `MrImputation` as a collection of R-functions `mrimpute` and `mranalyze`. *R*-function `mrimpute` performs multiple ratio imputation. *R*-function `mranalyze` allows us to conduct statistical analyses using the multiply-imputed data by *R*-function `mrimpute`. Takahashi (2017) offers a detailed explanation. As for single ratio imputation and multiple imputation, see de Waal et al. (2011), Hu et al. (2001), King et al. (2001), Carpenter & Kenward (2013), Honaker & King (2010), Honaker et al. (2011), Little & Rubin (2002).

Preparation Stage

As an illustration, consider the dataset `data`. In the code presented in Appendix, the name of `data` can be defined by option `data=`. Thus, it can be named any way an imputer wants it to be. This small example dataset contains two variables and five units as displayed in Figure 1. The observation for unit 1 in $y1$ is missing

Masayoshi Takahashi is an Assistant Professor of Institutional Research. Email at mtakahashi@tufs.ac.jp.

(NA). Thus, *y*1 is the target incomplete variable for imputation, and *y*2 is the auxiliary complete variable. Also, *y*1 is stored in `data[,1]` and *y*2 in `data[,2]`. This article will use this small dataset for illustration. As this dataset implies, the target variable for imputation needs to be stored in the first column of data, i.e., `data[,1]`, in order to execute the code shown in this article.

```
data<-read.csv("data.csv",header=T)
attach(data)
```

```
> data
        y1        y2
1       NA 10.545612
2 5.779933  9.728869
3 4.835343  9.920130
4 6.219675  8.897375
5 7.012357 10.417368
```

Figure 1. Example of Incomplete Data

The number of multiply-imputed data is set by `M`, where `M > 1`. In this example, it is set to 2 so that the outputs can be visually presented below. To allow reproducibility, the random number seed value needs to be set by function `set.seed`. This step is necessary, because multiple imputation relies on pseudo-random numbers; thus, without setting a seed, there will be no way of reproducing the same results.

```
M<-2
set.seed(1223)
```

Many types of data are skewed to the right in the distribution, i.e., the distribution is not multivariate normal, but multivariate log-normal. If this is the case, a sensible option to deal with such a variable is to use log-transformation, and the imputed values will be unlogged after imputations are completed (Allison, 2002, p.39; Honaker et al., 2011, p.15). In the complete code shown in Appendix, if `log=TRUE`, then the following code log-transforms the data. The default setting is that `log=FALSE`. Obviously, if data are multivariate normal to begin with, this option should be set to `FALSE`.

```
if(log){
  data<-log(data)
}
```

Imputation Stage

Nonparametric Bootstrap

The first step to perform multiple ratio imputation is to implement random draws of $\boldsymbol{\mu}$ from an appropriate posterior distribution to account for estimation uncertainty. The EMB algorithm substitutes the complex process of drawing $\boldsymbol{\mu}$ from the posterior distribution with a nonparametric bootstrapping algorithm, which is a resampling method, where the observed sample is used as the pseudo-population. In other words, a resample of size n is randomly drawn from this observed sample of size n with replacement, and this process is repeated M times (Shao & Tu, 1995; Horowitz, 2001).

R-function `sample(x,size,replace=TRUE)` can be used for this purpose, where `x` is a vector from which to sample, `size` is the number of items to sample, and `replace=TRUE` specifies sampling with replacement. Unfortunately, this function randomly draws a vector, not a matrix. In the process of imputation, the imputer must keep a pair of observations for the two variables. Thus, our code first creates `sampleframe` to randomly draw the row number of data, which is an `nrow(data)` by `M` matrix, where `nrow(data)` is the number of rows in `data`.

```
sampleframe<-matrix(sample(nrow(data),nrow(data)*M,replace=TRUE),
nrow=nrow(data),ncol=M)
```

The resulting matrix obtained from the above code is displayed in Figure 2, where each column contains a vector of the row numbers randomly drawn from the original data. For example, `sampleframe[1,1]` is 4, meaning that this cell refers to row number 4 in the original data, i.e., $y1 = 6.219675$ and $y2 = 8.897375$, `sampleframe[2,1]` is 1, meaning that this cell refers to row number 1 in the original data, i.e., $y1 =$ NA and $y2 = 10.545612$, and so on.

Based on `sampleframe`, our code makes a random draw of the values of $y1$ and $y2$ from the original data M times. First, let us create a list named `datasub` with the elements of NA and then replace these NAs by appropriate values in the original data, so that `datasub[[i]]` obtains `data[sampleframe[,i],]`, and the `for`

loop repeats this process *M* times. In order to use this `datasub` in the EM algorithm below, `datasub` is transformed to a matrix.

```
> sampleframe
     [,1] [,2]
[1,]    4    5
[2,]    1    1
[3,]    2    4
[4,]    2    5
[5,]    1    1
```

Figure 2. Randomly-Drawn Row Numbers

```
datasub<-as.list(rep(NA,M))
for(i in 1:M){
  datasub[[i]]<-as.matrix(data[sampleframe[,i],])
}
```

The resulting bootstrap resamples are shown in Figure 3, where `datasub[[1]]` and `datasub[[2]]` represent the m^{th} bootstrap resample, respectively.

```
> datasub
[[1]]
          y1        y2
4   6.219675  8.897375
1         NA 10.545612
2   5.779933  9.728869
2.1 5.779933  9.728869
1.1       NA 10.545612

[[2]]
          y1        y2
5   7.012357 10.417368
1         NA 10.545612
4   6.219675  8.897375
5.1 7.012357 10.417368
1.1       NA 10.545612
```

Figure 3. Example of Bootstrap Resamples (*M* = 2)

EM Algorithm

Each bootstrap resample created above is likely to be incomplete. Estimates using these resamples are expected to be biased and inefficient. In order to avoid this problem, the EM algorithm is used to refine the estimates in bootstrap resamples. As for the EM algorithm, see Little & Rubin (2002), Do & Batzoglou (2008), the *R*-package `Norm` by Schafer (1997), and the function `em.norm` (Fox, 2015). The current code does not use `Norm` for the sake of generating multiple imputation, but function `em.norm` is useful for the computational purpose of the EM algorithm. First, use the require function to load `Norm` in *R*. In the code below, `p` is the number of columns (variables) in the data, `para` is the number of parameters to be estimated, `thetahat` is an empty matrix with the dimension of `M` by `para`, and `emmu` is an empty matrix with the dimension of `M` by `p`. These are housekeeping issues to perform the EM algorithm by way of function `em.norm`.

```
require(norm)
p<-ncol(data)
para<-p*(p+3)/2+1
thetahat<-matrix(NA,M,para)
emmu<-matrix(NA,M,p)
```

Function `prelim.norm` takes care of the preliminary manipulations for a matrix of incomplete data, which is a necessary step for using `em.norm`, whose results are stored in `thetahat`. Option `showits=FALSE` quietly runs `em.norm`. If the imputer wants to monitor the iteration process of EM, then this option should be set to `TRUE`. Option `maxits=1000` sets the maximum number of iterations to 1,000. Function `getparam.norm` produces the estimated values of the MLEs, which is stored in `emmu`. Option `corr=FALSE` computes the means and variance-covariance matrix. The `for` loop repeats the `em.norm` function to be applied to `datasub` *M* times. This process is the essential part of the EMB algorithm, meaning that the EM algorithm is applied to each of the *M* bootstrap resamples.

```
for(i in 1:M){
  thetahat[i,]<-em.norm(prelim.norm(datasub[[i]]),
  showits=FALSE,maxits=1000)
  emmu[i,]<-getparam.norm(prelim.norm(datasub[[i]]),
  thetahat[i,],corr=FALSE)$mu
}
```

All of the estimates of the means by the EM algorithm are stored in `emmu`. Thus, typing `emmu` returns the following matrix in Figure 4, where the first column refers to the means for the first variable in the data, and the second column refers to the means for the second variable in the data. Also, the first row refers to the means in $m = 1$ and the second row refers to the means in $m = 2$. Note that these are the MLEs of the means.

```
> emmu
         [,1]      [,2]
[1,] 5.695139  9.889267
[2,] 6.880546 10.164667
```

Figure 4. MLEs for the Means of *y*1 and *y*2

Implementation of Multiple Ratio Imputation

Using matrix `emmu` allows us to estimate multiple ratios of two variables as follows. The estimated ratios are stored in `beta`, which is an empty matrix with the dimension of `M` by `ncol(data)-1`. Ratio imputation has only two variables; thus, the number of columns in the data, i.e., `ncol(data)`, is 2, which means that `beta` is essentially an `M` by 1 column vector.

```
beta<-matrix(NA,M,ncol(data)-1)
beta<-emmu[,1]/emmu[,2]
```

Typing `beta` returns a vector of M values, where the first value is the ratio in the first model, the second value in the second model, and so on. This is $\tilde{\beta}$ in equation (6) of Takahashi (2017).

```
> beta
[1] 0.5758909 0.6769082
```

Figure 5. The Values of the Slopes in the Multiple Ratio Imputation Model

As a preparation for multiple ratio imputation, let us define the following matrices. These are housekeeping issues to perform multiple ratio imputation. All of the matrices are empty matrices with the dimensions of `nrow(data)` by `M`.

```
imp<-matrix(NA,nrow(data),M)
resid<-matrix(NA,nrow(data),M)
e<-matrix(NA,nrow(data),M)
imp1<-matrix(NA,nrow(data),M)
imp2<-matrix(NA,nrow(data),M)
```

The values of `beta` are multiplied by `data[,2]` which is the values of the second variable in the data. Specifically, `data[,2]` is *y*2 in our example. Thus, the following code is $\tilde{\beta}Y_{i2}$ in equation (6) of Takahashi (2017). The `for` loop repeats this process *M* times. The imputed values are stored in `imp`, where `imp[,1]` is the imputed data from `m = 1` and `imp[,2]` is the imputed data from `m = 2`.

```
for(i in 1:M){
  imp[,i]<-beta[i]*data[,2]
}
```

To complete the process, a small disturbance term needs to be added to the imputed values, which is $\tilde{\varepsilon}_i$ in equation (6) of Takahashi (2017). In the following code, `resid` is the differences (residuals) between observed values and predicted values. Also, $\tilde{\varepsilon}_i$ is `e[,i]`, which is normally distributed with the mean of 0 and the standard deviation of the residuals, `resid[,i]`. In the last line, `e[,i]` is added to `imp[,i]`. The `for` loop repeats this whole process *M* times.

```
for(i in 1:M){
  resid[,i]<-data[,1]-imp[,i]
  e[,i]<-rnorm(nrow(data),0,sd(resid[,i],na.rm=TRUE))
  imp1[,i]<-imp[,i]+e[,i]
}
```

All of the values were imputed, both observed and missing. What actually needs to be imputed is the missing part of the data only. Therefore, the final step is to replace `NA` with `imp1` and to keep the observed value as is. In the following code, `imp2` is essentially $\tilde{Y}_{i1}$ in equation (6) of Takahashi (2017). If `data[j,1]` is

missing, then `imp2[j,i]` obtains the imputed value `imp1[j,i]`; otherwise, `imp2[j,i]` obtains `data[j,1]`. In the following loop, `i` refers to the number of imputations and `j` refers to the row number in the data.

```
for(i in 1:M){
  for(j in 1:nrow(data)){
    if (is.na(data[j,1])=="TRUE"){
    imp2[j,i]<-imp1[j,i]
  }else{
    imp2[j,i]<-data[j,1]}
}}
```

Remember that log-normal data were log-transformed above. Imputed values must be put back to the original scale of incomplete data. The following code unlogs the log-transformed variables.

```
if(log){
  imp2<-exp(imp2)
  data<-exp(data)
}
```

Some variables have logical bounds. For instance, economic variables such as turnover cannot be negative. If this is the case, `zero=TRUE` can be specified in the complete code in Appendix. This option forces negative imputed values to be zero. Warning is that this option may suppress the correct uncertainty in the imputation model (Honaker et al., 2011, pp. 23-25); thus, this option should be used cautiously. The default setting is `zero=FALSE`.

```
if(zero){
  imp2[which(imp2<0)]<-0
}
```

Finally, `imp2` returns the following two sets of imputed data, because $M = 2$. The values in row [1,] change over columns [,1] to [,2], because these values are imputed values. The values in the other rows do no change over columns, because these are observed values.

```
> imp2
         [,1]     [,2]
[1,] 6.739130 6.828206
```

```
[2,] 5.779933 5.779933
[3,] 4.835343 4.835343
[4,] 6.219675 6.219675
[5,] 7.012357 7.012357
```

Figure 6. Example of Multiply-Imputed Data

The `write.csv` function saves the imputed data along with the original data as follows, where `y1` is the original incomplete variable, `y2` is the original auxiliary variable, and `imp2` is a matrix of *M* imputed data created above.

```
y1<-data[,1]; y2<-data[,2]
impdata<-data.frame(y1,y2,imp2)
write.csv(impdata,"mridata.csv",row.names=FALSE)
```

Figure 7 contains the output data named `mridata` in the csv format, which can be reloaded in *R* or any statistical software of an analyst's choice for subsequent statistical analyses. In this output dataset, Column A (*y*1) is the original incomplete data, Column B (*y*2) is the original auxiliary variable, and Columns C to D (*X*1, *X*2) are the multiply imputed data.

A10

	A	B	C	D
1	y1	y2	X1	X2
2	NA	10.54561	6.73913	6.828206
3	5.779933	9.728869	5.779933	5.779933
4	4.835343	9.92013	4.835343	4.835343
5	6.219675	8.897375	6.219675	6.219675
6	7.012357	10.41737	7.012357	7.012357

Figure 7. Example of Output Data (csv file)

Analysis Stage

Mean and Standard Deviation

After reading `mridata.csv`, various statistical analyses can be performed. To calculate the mean and the standard deviation of an imputed variable ($y1$), the analyst first creates two empty vectors of `means` and `sds`, and repeats the calculations M times by the `for` loop. Typing `means` and `sds` returns M values of the means and the standard deviations.

```
means<-c(NA); sds<-c(NA)
  for(k in 1:M){
  means[k]<-mean(imp2[,k])
  sds[k]<-sd(imp2[,k])
}
```

To calculate a combined point estimate, the analyst simply takes the average by equation (7) of Takahashi (2017). Furthermore, by calculating the standard deviation of `means`, i.e. `sd(means)`, the analyst can estimate the amount of estimation uncertainty due to imputation as a confidence interval.

```
mean(means)                 #Combined Point Estimate of Mean
mean(sds)                   #Combined Point Estimate of Std. Dev.
sd(means)                   #Estimation Uncertainty
mean(means)+2*sd(means)     #Confidence Interval Upper Limit
mean(means)-2*sd(means)     #Confidence Interval Lower Limit
```

Consider again the example data in Figure 1. The combined point estimate of the means is 6.126, with the combined point estimate of standard deviation 0.868. Estimation uncertainty is measured by `sd(means)`, which is the standard deviation of the M means, or the standard error of the estimated M means. In our case, it is 0.013. Therefore, there is an approximately 95% confidence that the true mean of complete data is somewhere between 6.101 and 6.151, after taking the error due to missingness into account.

Regression of $y2$ on $y1$

Suppose that $y2$ is the dependent variable and $y1$ is the explanatory variable in regression. To estimate the regression coefficients and the associated standard

errors, the analyst first creates four empty vectors, `reg1`, `reg2`, `reg3`, and `reg4`. The `for` loop repeats the estimation of regression models *M* times. The results are stored in `summary(model)$coefficients[i]`, where *i* = 1 and 3 are regression coefficients and *i* = 2 and 4 are standard errors.

```
reg1<-c(NA); reg2<-c(NA); reg3<-c(NA); reg4<-c(NA)
for(k in 1:M){
  model<-lm(data[,2]~data[,k+2])
  reg1[k]<-summary(model)$coefficients[1]
  reg2[k]<-summary(model)$coefficients[2]
  reg3[k]<-summary(model)$coefficients[3]
  reg4[k]<-summary(model)$coefficients[4]
}
```

After the analysis stage is complete, there are *M* values of outputs. Using equations (7) and (8) of Takahashi (2017), the results are combined as follows.

```
intercept<-mean(reg1)              #Combined Intercept
WV1<-mean(reg3^2)                  #Within-Imputation Variance
BV1<-sum((reg1-intercept)^2)/(M-1) #Between-Imputation Variance
TV1<-WV1+(1+1/(M))*BV1             #Total Variance
TSE1<-sqrt(TV1)                    #Total Std. Error
tstat1<-intercept/TSE1             #t-statistics for Intercept
slope<-mean(reg2)                  #Combined Slope
WV2<-mean(reg4^2)                  #Within-Imputation Variance
BV2<-sum((reg2-slope)^2)/(M-1)     #Between-Imputation Variance
TV2<-WV2+(1+1/(M))*BV2             #Total Variance
TSE2<-sqrt(TV2)                    #Total Std. Error
tstat2<-slope/TSE2                 #t-statistics for Slope
```

Consider again the example data in Figure 1. The combined point estimate of the regression intercept is 8.231, with the total standard error of 2.512. Thus, the *t*-statistic for the intercept is 3.277. The combined point estimate of the regression slopes is 0.273 with the total standard error of 0.407. Thus, the *t*-statistic for the slope is 0.671.

Conclusion

It was outlined here how to implement multiple ratio imputation in *R*, which can be easily copied and pasted into *R* for use (See Appendix). These codes estimate multiple ratio imputation, and statistically analyze imputed data by multiple ratio imputation. Therefore, this will be a valuable addition to the choice for imputation techniques.

However, the code described here is only a first step toward implementing multiple ratio imputation; thus, the code is expected to be updated so as to maximize computational efficiency and to expand the scope of data that can be handled. Furthermore, the EMB algorithm is a general approach composed of the EM algorithm and nonparametric bootstrapping. Therefore, multiple ratio imputation can be implemented not only in *R*, but also in other statistical environments. Also, multiple ratio imputation is not limited to the EMB algorithm. Depending on the nature of imputation, multiple ratio imputation may be implemented by way of other multiple imputation algorithms, such as MCMC and Fully Conditional Specification (FCS) (van Buuren, 2012).

Acknowledgments

The author wishes to thank Mr. Yutaka Abe (Hitotsubashi University) for his valuable comments on an earlier version of this article. The author also wishes to thank the two anonymous reviewers for reviewing this article. However, any remaining errors are the author's responsibility. Also, note that the views and opinions expressed in this article are the author's own, not necessarily those of the institution. The analyses in this article were conducted using *R*.3.1.0.

References

Allison, P. D. (2002). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple Imputation and its Application*. Chichester, West Sussex: John Wiley & Sons. doi: 10.1002/9781119942283

de Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470904848

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology, 26*(8), 897-899. doi: 10.1038/nbt1406

Fox, J. (2015). Package 'Norm' [Computer software]. Retrieved from: http://cran.r-project.org/web/packages/norm/norm.pdf

Honaker, J., & King, G. (2010). What to do about missing values in time series cross-section data. *American Journal of Political Science, 54*(2), 561-581. doi: 10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: a program for missing data. *Journal of Statistical Software, 45*(7), 1-47. doi: 10.18637/jss.v045.i07

Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman & E. Leamer (Eds), *Handbook of Econometrics* (pp. 3160-3228), Vol. 5. Amsterdam: Elsevier. doi: 10.1016/s1573-4412(01)05005-x

Hu, M., Salvucci, S., & Lee, R. (2001). *A Study of Imputation Algorithms. Working Paper No. 2001–17*. U.S. Department of Education. National Center for Education Statistics. Retrieved from: http://nces.ed.gov/pubs2001/200117.pdf

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review, 95*(1), 49-69.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781119013563

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.

Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap*. New York, NY: Springer. doi: 10.1007/978-1-4612-0795-5

Takahashi, M. (2017). Multiple ratio imputation by the EMB algorithm: Theory and simulation. *Journal of Modern Applied Statistical Methods*, 16(1). doi: 10.22237/jmasm/1493598840

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/b11826

Appendix: Software `MrImputation`

Software `MrImputation` (version 1.0.0), which stands for multiple ratio imputation, is a collection of *R*-functions explained step by step in this article. This appendix combines each of the steps as a set of *R*-functions `mrimpute` and `mranalyze`.

User Manual

Copy the following codes into the *R* script and save them as `mrimpute.R` and `mranalyze.R` on the computer. After reading an appropriate data file in *R*, use function source to read these functions as follows.

```
source("mrimpute.R")
source("mranalyze.R")
```

Description of `mrimpute` This function performs the imputation stage of multiple ratio imputation and produces multiply-imputed data named mridata.csv.

Usage `mrimpute(data = data, M = 100, seed = 1223, log = FALSE, zero = FALSE, outdata = TRUE)`

Arguments

`data`	A data frame that contains the incomplete variable targeted for imputation. The imputer can specify any name of the data to be used.
`M`	The number of multiply-imputed datasets. The imputer can set any number.
`seed`	Random number seed value. Any number can be specified.
`log`	An option to log-transform the data. The default is FALSE. If log-transformation is optimal, then this option should be set to TRUE.
`zero`	An option to suppress negative values to zero. The default is FALSE. If negative imputed values are unacceptable, this option should be set to TRUE.
`outdata`	An option to save the imputed data as a csv file. The default is TRUE.

Description of `mranalyze` This function performs the analysis stage. It returns the mean and the standard deviation of the imputed variable. It can also return the result of regression analysis of $y2$ on $y1$ if `reg=TRUE`.

Usage `mranalyze(data, reg = FALSE)`

Arguments

`data`	The mridata.csv created by mrimpute.
`reg`	An option to perform regression analysis. The default is FALSE. If the analyst wants to see the result of regression analysis, this option should be set to TRUE.

R-Function `mrimpute`: Imputation Stage

```
mrimpute<-function(data,M,seed,outdata=TRUE,log=FALSE,zero=FALSE){
data<-data; M<-M; seed<-seed; set.seed(seed)
if(log){data<-log(data)}
sampleframe<-matrix(sample(nrow(data),nrow(data)*M,
             replace=TRUE),nrow=nrow(data),ncol=M)
datasub<-as.list(rep(NA,M))
for(i in 1:M){datasub[[i]]<-as.matrix(data[sampleframe[,i],])}
suppressMessages(suppressWarnings(require(norm)))
p<-ncol(data); para<-p*(p+3)/2+1; thetahat<-matrix(NA,M,para)
emmu<-matrix(NA,M,p)
for(i in 1:M){thetahat[i,]<-em.norm(prelim.norm(datasub[[i]]),
                           showits=FALSE,maxits=1000)
              emmu[i,]<-getparam.norm(prelim.norm(datasub[[i]]),
                           thetahat[i,],corr=FALSE)$mu}
imp0<-as.list(rep(NA,M)); imp<-matrix(NA,nrow(data),M)
resid<-matrix(NA,nrow(data),M); e<-matrix(NA,nrow(data),M)
imp1<-matrix(NA,nrow(data),M); beta<-matrix(NA,M,ncol(data)-1)
beta<-emmu[,1]/emmu[,2]
for(i in 1:M){imp[,i]<-beta[i]*data[,2]}
for(i in 1:M){resid[,i]<-data[,1]-imp[,i]
              e[,i]<-rnorm(nrow(data),0,sd(resid[,i],na.rm=TRUE))
              imp1[,i]<-imp[,i]+e[,i]}
imp2<-matrix(NA,nrow(data),M)
for(i in 1:M){imp2[,i]<-data[,1]}
for(i in 1:M){
  for(j in 1:nrow(data)){
    if (is.na(data[j,1])=="TRUE"){
```

```
    imp2[j,i]<-imp1[j,i]
    }else{
    imp2[j,i]<-data[j,1]}
  }}
if(log){imp2<-exp(imp2);data<-exp(data)}
if(zero){imp2[which(imp2<0)]<-0}
impdata<-data.frame(data, imp2)
  if (outdata){
    write.csv(impdata,"mridata.csv",row.names=FALSE)
  }
}
```

R-Function `mranalyze`: Analysis Stage

```
mranalyze<-function(data,reg=FALSE){
data<-data; M<-ncol(data)-2; means<-c(NA); sds<-c(NA)

for(k in 1:M){
  means[k]<-mean(data[,k+2])
  sds[k]<-sd(data[,k+2])
}
meanimp<-mean(means);BISD<-sd(means);UL<-mean(means)+2*sd(means);LL<-
           mean(means)-2*sd(means);sd<-mean(sds)
outmatrix1<-matrix(c(meanimp, sd, BISD, UL, LL))
colnames(outmatrix1)<-"Summary"
rownames(outmatrix1)<-c("mean","sd","BISD","95%CIUL","95%CILL")

if(reg){
reg1<-c(NA); reg2<-c(NA); reg3<-c(NA); reg4<-c(NA)
for(k in 1:M){
  model<-lm(data[,2]~data[,k+2])
  reg1[k]<-summary(model)$coefficients[1]
  reg2[k]<-summary(model)$coefficients[2]
  reg3[k]<-summary(model)$coefficients[3]
  reg4[k]<-summary(model)$coefficients[4]
}

intercept<-mean(reg1)
```

```
WV1<-mean(reg3^2)
BV1<-sum((reg1-intercept)^2)/(M-1)
TV1<-WV1+(1+1/(M))*BV1
TSE1<-sqrt(TV1)
tstat1<-intercept/TSE1

slope<-mean(reg2)
WV2<-mean(reg4^2)
BV2<-sum((reg2-slope)^2)/(M-1)
TV2<-WV2+(1+1/(M))*BV2
TSE2<-sqrt(TV2)
tstat2<-slope/TSE2

outmatrix2<-matrix(c(intercept, TSE1, tstat1, slope, TSE2, tstat2))
colnames(outmatrix2)<-"Regression"
rownames(outmatrix2)<-c("intercept","TSE(intercept)","t-
           Stat(intercept)","slope","TSE(slope)" ,"t-Stat(slope)")
}

if(reg){
result<-list(outmatrix1, outmatrix2)
  return(result)
}else{
result<-list(outmatrix1)
  return(result)
}
}
```

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 674-688.
doi: 10.22237/jmasm/1493598960

ISSN 1538 − 9472

An Unbiased Estimator Of The Greatest Lower Bound

Nol Bendermacher
Radboud University Nijmegen
Nijmegen, Netherlands

The greatest lower bound to the reliability of a test, based on a single administration, is the Greatest Lower Bound (GLB). However the estimate is seriously biased. An algorithm is described that corrects this bias.

Keywords: test reliability, greatest lower bound, GLB, unbiased estimate, capitalization on chance

Introduction

In classical test theory the concept of reliability refers to the precision of a measurement. In order to estimate the reliability of a test one needs two or more measurements applied to the same subjects. However, in many situations it is impossible to repeat a test administration under the same conditions. The next best thing is to estimate a lower bound to the reliability.

The current study is restricted to the reliability of tests that consist of a number of items and to the situation where the test is administered only once. The total score is the sum of scores on the individual items. According to classical test theory, the score x_{ij} of person i on item j consists of two parts: the true score τ_{ij} and an error component ε_{ij} : $x_{ij} = \tau_{ij} + \varepsilon_{ij}$. The error component includes not only real measurement errors but also the information that is unique to the item. It is assumed that these error components are uncorrelated with the true parts, as well as with each other. As a consequence the covariance matrix $\mathbf{\Gamma}$ of the items is the sum of two component matrices: the covariance matrix $\mathbf{\Gamma}_\tau$ of the true parts and the covariance matrix $\mathbf{\Gamma}_\varepsilon$ of the error components:

Nol Bendermacher is a retired member of the Research Technische OndersteuningsGroep (research technical support group). Email at bendermacher@hotmail.com.

$$\mathbf{\Gamma} = \mathbf{\Gamma}_\tau + \mathbf{\Gamma}_\varepsilon$$

The assumption of uncorrelated errors implies that $\mathbf{\Gamma}_\varepsilon$ is a diagonal matrix. Therefore the off-diagonal cells of $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_\tau$ are identical.

The reliability of a test consisting of v items is defined as:

$$\rho_{tt} = 1 - \frac{\sigma_e^2}{\sigma_a^2} \tag{1}$$

where σ_e^2 is the sum of the error variances, and σ_a^2 is the variance of the test scores.

$$\sigma_e^2 = TR\left(\mathbf{\Gamma}_e\right) = \sum_{i=1}^{v} \mathbf{\Gamma}_{eii} \tag{2}$$

$$\sigma_a^2 = \sum_{i=1}^{v}\sum_{j=1}^{v} \mathbf{\Gamma}_{ij} \tag{3}$$

According to these definitions the formula of reliability can be rewritten as

$$\rho_{tt} = 1 - \frac{\sigma_e^2}{\sigma_a^2} = 1 - \frac{\sum_{i=1}^{v} \mathbf{\Gamma}_{eii}}{\sum_{i=1}^{v}\sum_{j=1}^{v} \mathbf{\Gamma}_{ij}} \tag{4}$$

The Greatest Lower Bound

From (4) it becomes clear, given the covariance matrix $\mathbf{\Gamma}$, that the reliability is maximal if the trace of the error covariance matrix $\mathbf{\Gamma}_e$ is minimal. As Jackson and Agunwamba (1977) remark, the only restrictions that the classical model imposes on the elements of $\mathbf{\Gamma}_\varepsilon$ are

(1) $0 \leq \mathbf{\Gamma}_{eii} \leq \mathbf{\Gamma}_{ii}$ (5)

(2) $\mathbf{\Gamma}_\tau = \mathbf{\Gamma} - \mathbf{\Gamma}_\varepsilon$ is non-negative definite

Therefore, if the set of values $\mathbf{\Gamma}_e$ can be found that maximizes its trace under these restrictions, the result is the smallest possible value for the reliability, given the covariance matrix $\mathbf{\Gamma}$. This value is the greatest possible lower bound to the reliability, called the GLB. Its possible values are restricted to the range [0,1]. A procedure to estimate it from a given covariance matrix is described in Ten Berge, Snijders and Zegers (1981).

A serious problem with the GLB is that it suffers from a phenomenon known as capitalization on chance: if it is estimated from a sample it tends to overestimate the population value. The bias increases with decreasing sample size and with lower values of the GLB; see Shapiro and ten Berge (2000). Moreover, the bias will be larger with a larger number of items.

To illustrate the seriousness of the problem: imagine a set of 40 items, completely uncorrelated and all with a unit normal distribution. Because the covariance matrix of these items is diagonal, the GLB for the test is zero. However, if samples of size 200 are drawn from the population, the average GLB-estimate from these samples is about 0.56.

Finding an unbiased estimator

Bendermacher (2010) describes an algorithm which reduces the bias in the estimated GLB by the use of a bootstrapping procedure. A large number of samples are drawn (with replacement) from the observed data with sample sizes equal to the size of the observed sample. For each sample the GLB is computed and the difference between the average of the sample-GLBs and the observed GLB is taken as an estimate of the bias. If this difference is subtracted from the observed GLB, the result is a less-biased estimate. The algorithm to be explained in this article starts in the same way, but it proceeds a few steps further and thereby manages to reduce the bias to a negligible quantity.

The algorithm tries to reconstruct the population covariance matrix $\mathbf{\Gamma}$ and then takes the GLB of this reconstructed matrix $\mathbf{G}_p$ as an unbiased estimator of the population GLB. The reconstruction is based on the following simple starting points:

1. The population-GLB β is smaller than the observed sample-GLB b_o. Theoretically this is incorrect (take for instance the case $\beta = 1$), but in almost all practical situations it will hold.

2. The population matrix $\mathbf{\Gamma}$ is similar to the sample covariance matrix $\mathbf{G}_o$.

3. If samples $\mathbf{G}_s$ are drawn from the reconstructed covariance matrix $\mathbf{G}_p$ (with the same size as the sample from which the observed matrix $\mathbf{G}_o$ was computed) their uncorrected GLB has as its expectation the observed value b_o.

The reconstruction of $\mathbf{\Gamma}$ will be called $\mathbf{G}_p$. It is built by adjustments to $\mathbf{G}_o$, which lower the value of its GLB. Because the three starting points still leave a considerable room in the exact way they are operationalized, several approaches were investigated, like adding error variances to the diagonal of $\mathbf{G}_o$, shrinking the off-diagonal cells, and reflecting some items to make their item-rest correlations negative. All these methods succeed in finding a covariance matrix that complies with the three starting points, but that does not mean by itself that the resulting GLB is an unbiased estimator. After some trial and error based on analyses of samples from two large real life data files, the following procedure appears to produce the best results by far:

1. Given the observed covariance matrix $\mathbf{G}_o$, compute the estimate $\mathbf{G}_t$ of $\mathbf{\Gamma}_\tau$ with on its diagonal the minimal true variances and with its off-diagonals equal to those of $\mathbf{G}_o$. Example:

$\mathbf{G}_o$ =				$\mathbf{G}_t$ =			
6.4259	…	…	…	3.0717	…	…	…
3.0040	3.9210	…	…	3.0040	3.6019	…	…
1.5511	1.2191	5.0580	…	1.5511	1.2191	0.9501	…
1.2958	0.3373	1.0951	14.3406	1.2958	0.3373	1.0951	1.8588

The GLB of $\mathbf{G}_o$ is $b_o = 0.5666$.

2. Multiply the diagonals of $\mathbf{G}_t$ by a factor $c \leq 1$. Call the resulting matrix $\mathbf{G}^*$.The rationale is that if $\mathbf{\Gamma}$ has a lower GLB than $\mathbf{G}_o$ its minimal true variances must be relatively smaller. How the factor c should be chosen will be explained later on.

   The example with $c = 0.69543$:

$\mathbf{G}^*$ =

2.1358	…	…	…
3.0040	2.5045	…	…
1.5511	1.2191	0.6606	…
1.2958	0.3373	1.0951	1.2924

3. Due to its lowered diagonal elements, $\mathbf{G}^*$ will have some negative eigenvalues.

   Compute the eigenvectors $\boldsymbol{V}$ and eigenvalues Λ of $\mathbf{G}^*$, such that $\mathbf{G}^* = \boldsymbol{V}\Lambda\boldsymbol{V}^T$. Example:

Λ =

6.4570	…	…	…
…	1.4324	…	…
…	…	-0.3546	…
…	…	…	-0.9415

4. Replace the negative eigenvalues of $\mathbf{G}^*$ by zeros and add their (negative) values to the smallest non-negative eigenvalues without letting them become negative. Call the result Λ^*. Example:

Λ^* =

6.4570	…	…	…
…	1.3630	…	…
…	…	0.0000	…
…	…	…	0.0000

5. Compute $\mathbf{G}^* = \boldsymbol{V}\Lambda^*\boldsymbol{V}^T$; its trace will be $c.TR(\mathbf{G}_t)$

6. Complete the reconstruction of the population matrix by replacing the diagonal of $\mathbf{G}^*$ by that of $\mathbf{G}_o$: $\mathbf{G}_p = \mathbf{G}^* - \text{DIAG}(\mathbf{G}^*) + \text{DIAG}(\mathbf{G}_o)$. Example:

$\mathbf{G}_p$ =

6.4259	…	…	…
2.5760	3.9210	…	…
1.4686	1.4016	5.0580	…
1.1435	1.0547	0.6671	14.3406

7. Compute the GLB b_p of $\mathbf{G}_p$. This is the corrected estimate of the population GLB. In the example, $b_p = 0.5005$.

There remains a crucial question: what is the correct value of the factor c in step 2 of the above procedure. The answer is based on the third starting point. The factor c must be chosen such that the expected GLB of samples from $\mathbf{G}_p$ is equal to the observed GLB b_o. This means that one can start from a well chosen guess c, compute $\mathbf{G}_p$ and perform a bootstrapping run in which a large number of samples matrices $\mathbf{G}_{si}$ are drawn from $\mathbf{G}_p$.

The average b_s of the sample GLB-values, as compared to the observed GLB b_o, is used to update the choice of c, and the process is repeated until the correct value has been found. More details are given in the section Algorithm. This procedure requires several bootstrapping runs, each generating a vast number of samples. Therefore it is important to have an efficient algorithm that keeps the number of bootstrap runs at a minimum.

Drawing samples from a covariance matrix

How a sample covariance matrix can be derived from a population matrix without knowing the underlying raw data will now be explicated. The algorithm requires covariance matrices based on samples from the data from which $\mathbf{G}_o$ is computed. If these data are available one might actually draw such samples and compute covariance matrices from them. However, because the algorithm implies a number of bootstrapping runs, with a large amount of samples for each run, such a procedure would be very time consuming. Moreover, the algorithm also requires sampling from modified covariance matrices for which no raw data are available. Fortunately it is possible to compute these sample covariance matrices directly from the observed or constructed covariance matrix and the given or assumed distributions of the items.

If a sample of raw data is given, estimates of the distributions of the items can be derived from that sample. If no information is available about the distributions of the items one may assume a multivariate normal distribution.

Sampling from a given $v \times v$ covariance matrix $\mathbf{G}$ with sample size k can be performed as follows:

1. Compute, by Cholesky triangularization, a matrix $\mathbf{C}$ such that $\mathbf{C}\mathbf{C}^T = \mathbf{G}$.

2. Generate k times a vector of v independently chosen random drawings using the distributions of the v items. Compute the covariance matrix $\mathbf{G}_z$ from these vectors, as if they were observed cases.

3. Compute a matrix $\mathbf{G}^*$ by dividing each cell of $\mathbf{G}_r$ by the standard deviations of the two items involved: $\mathbf{G}^*_{ij} = \dfrac{\mathbf{G}_{zij}}{s_i s_j}$.

4. Compute the sample matrix as $\mathbf{G}_s = C\mathbf{G}^*C^T$

The average of the GLB-values of the matrices $\mathbf{G}_z$ (see step 2) gives an estimate of the expected sample GLB b_z under the null hypothesis that $\mathbf{G}_p$ has GLB-value zero. If the observed GLB (b_o) is clearly less than b_z the corrected estimate b_p can immediately be set to zero.

If one assumes a multivariate normal distribution of the items, the v independently chosen drawings mentioned in step 2 can be drawings from a unit normal distribution. To speed up the program one may construct in advance a long list (say 4000 numbers) of drawings from a unit normal distribution by taking equally spaced values between 0 and 1 and computing the inverse of the cumulative normal distribution function for them. Sampling from a unit normal distribution then comes down to randomly choosing from this list, using a uniform random generator.

Algorithm

This description of the algorithm uses the following definitions:

`Go`	the observed covariance matrix
`Gp`	the current reconstruction of $\mathbf{\Gamma}$
`bo`	the GLB of `Go`
`bp`	the GLB of `Gp`, i.e. the provisional estimate of β
`bs`	the average GLB of the samples from the most recent bootstrap run
`bz`	the average of the GLB-values of samples simulated under the null hypothesis of uncorrelated items
`bt`	the intended GLB-value for an updated reconstruction `Gp`

The algorithm consists of the following steps:

Step 1: Choose a precision criterion `Precision`; 0.001 will do well.

Choose `MaxSteps` = the maximum number of steps in the main algorithm; suggested value: 100.

Set `CurrentPrecision` = `Precision` × 5; set `ShrinkFactor` = $0.2^{1/5}$

`ShrinkFactor` will be used to decrease `CurrentPrecision` in five steps towards `Precision`.

Step 2: Perform a bootstrap run in which samples are drawn from `Go` until the standard error of the mean of sample GLB-values is less than `CurrentPrecision` or a maximum number of samples is drawn.

The main results are: *bz*, *bs* and `Significance`. `Significance` gives the proportion of samples generated under the null hypothesis of uncorrelated items with a GLB-value greater than *bo*.

Step 3: If *bo* < *bz* × 0.9 or `Significance` ≥ 0.5, then set `Bestbp` = 0 and go to step 16

Step 4: Initialize some variables: `BestDiff` = 9, `Bestbp` = *bo*, `BestCount` = 0, `Count` = 0

Step 5: Find successive new versions of the reconstructed population matrix `Gp` by repeating steps 6-15

Step 6: Increase `Count`; If `Count` > `MaxSteps` go to step 16

Step 7: Find a new *bt*:

If *bs* ≤ *bo* then

set `LowLim` = MIN(*bs*,*bp*)

set `UppLim` = MAX(`LowLim`,`UppLim`)

set *bt* = (`LowLim` + `UppLim`) / 2

else perform steps 7a - 7d

Step 7a: Set `UppLim` = *bp*

Step 7b: Set LowLim = MIN(LowLim,UppLim)

Step 7c. Find a second order polynomial $y = f(x)$ through the points $(x,y) = (bz,0)$, (bs,bp) and $(1,1)$

and find $bt = f(bo)$.

Compute the predictor matrix **P** and the criterion vector ***Q***:

$$\mathbf{P} = \begin{bmatrix} b_z^2 & b_z & 1 \\ b_s^2 & b_s & 1 \\ 1 & 1 & 1 \end{bmatrix}; \boldsymbol{Q} = \begin{bmatrix} 0 \\ b_p \\ 1 \end{bmatrix}$$

If **P** is singular set
bt = MIN(1,MAX(0,bp − (bs − bo) × 1.2)

else compute the weights $W = \mathbf{P}^{-1}\boldsymbol{Q}$ and set
$bt = W_1 b_o^2 + W_2 b_o + W_3$

Step 8. IF Count = 1 set bt= MIN(bt,0.95)

Step 9. Find a new estimate Gp such that its GLB bp is close enough to bt, i.e. until ABS(bp−bt) < CurrentPrecision) or a maximum of steps is taken.

Compute the GLB bp of Gp. The details of this step are described later.

Step 10. Perform a bootstrap run and compute the average value bs of the sample GLB's.

Step 11. Compute Diff = ABS(bs−bo)

If Diff < BestDiff then

set BestDiff = Diff; set Bestbp = bp; and set BestCount = Count

Step 12. If Diff ≤ CurrentPrecision then

If CurrentPrecision = Precision go to step 16

else set CurrentPrecision = Precision

Step 13. If BestCount $\leq$ Count $-$ 5 then

If CurrentPrecision = Precision go to step 16

else set CurrentPrecision = Precision

Step 14. If Count < 4
set CurrentPrecision = ShrinkFactor $\times$ CurrentPrecision

If Count = 4 set CurrentPrecision = Precision

Step 15. Go back to step 6

Step 16. Set *bp* = MAX(0,MIN(Bestbp,1)

Step 17. Now *bp* is the final value of the corrected GLB

Some explanations:

at Step 1: The algorithm may be very time consuming. Therefore the required precision is varied from 5 times Precision in the first cycle to Precision in the fifth and following cycles.

at Step 9: The factor *c* and the corresponding matrix G*p* can be found by the following algorithm:

Step 9a. Set Lowc = 0; Set Highc = 1; set Lowb = *bz*; set Highb = *bo*

Step 9b. Repeat steps 9c through 9h

Step 9c. Set Midc = (Lowc + Highc)/2

If ABS(Highb $-$ Lowb) < CurrentPrecision go to step 9i

Step 9d. Copy G*o* to G*p*

Step 9e. If MidC $\geq$ 1 $-$ Precision set Midb = *bo*

else ... (steps 9f through 9h)

Step 9f. Replace the diagonal of Gp by Midc times the vector of minimal true variances of Go

Compute the eigenvectors $\boldsymbol{V}$ and the diagonal matrix $\boldsymbol{\Lambda}$ with eigenvalues of Gp

Step 9g. Set $T_1 = TR(Gp)$; set T_2 = sum of the negative eigenvalues in $\boldsymbol{\Lambda}$.

Replace the negative eigenvalues by zero.

Loop over the positive eigenvalues λ_i from smallest to greatest:

If $\boldsymbol{\Lambda}_{i,i} \geq T_2$ then set $\boldsymbol{\Lambda}_{i,i} = \boldsymbol{\Lambda}_{i,i} - T_2$ and continue with step 9h

else set $T_2 = T_2 - \boldsymbol{\Lambda}_{i,i}$ and set $T_2 = 0$; continue the loop over the eigenvalues

Step 9h. Recompute $Gp = \boldsymbol{V\Lambda V}$ with the adjusted eigenvalues given by $\boldsymbol{\Lambda}$

Replace the diagonal of Gp by that of Go and compute its GLB *bp*.

Set Midb = *bp*

Step 9i. If ABS(*bt* − Midb) < CurrentPrecision go to step 9k

If a maximum (e.g. 30) number of cycles (9c through 9h) is taken go to step 9k

If *bt* < MidB set HighC = MidC

else set LowC = MidC

Step 9j. Go back to step 9c

Step 9k. Now `Gp` is the wanted matrix with its GLB `bp` close to `bt`.

Border effects

The correction procedure as it was specified above may fail for extreme observed GLB-values b_o. For low values, there may be no population matrix possible with b_o as its expected sample value. This happens if the observed GLB is lower than the expected sample value b_z under the null hypothesis $b_p = 0$. In such cases the corrected estimate can immediately be set to 0. For high values of b_o, the problem is not that easy to be solved. If the observed GLB b_o is (almost) 1, the estimator $b_p = 0.99...$ complies with the three starting points, but samples from a population with a lower value might as well have a GLB equal to or close to 1. In such cases the algorithm may erroneously overestimate the population GLB.

Evaluating the estimation procedure

In order to test the quality of the above procedure several large datasets were downloaded (personality-testing.info, n.d.), not including the files used in the trial and error phase. From each of these datasets one or more tests were selected and from each test 100 or 50 samples were taken, consisting of randomly chosen cases. Cases with missing values were not allowed to enter the samples.

As a result several sets were available each consisting of a large population and 100 or 50 samples extracted from it. The mean of the corrected GLB-values over the samples renders an estimate of the expected value of the corrected GLB. If the correction algorithm works correctly, these expected corrected GLB's should be (almost) equal to their corresponding population values. The tests were taken from the following data collections:

1. 16PF, test 1, items A1-A10, ordinal scores (1-5), 49159 cases
2. 16PF, test 2, items B1-B13, ordinal scores (1-5), 49159 cases
3. 16PF, test 3, items C1-C10, ordinal scores (1-5), 49159 cases
4. ECR, items Q1-Q36, ordinal scores (1-5), 17386 cases
5. MSSCQ, items Q1-Q100, ordinal scores (1-5), 17685 cases

Table 1 summarizes the main results, with column definitions as follows:

test	name of the test
# files	number of sample files taken from the large population file
v	test length
n	sample size
β	the GLB-value computed from the large population file; the average b_p (in column 7) should be close to this value
b_o	the mean of the uncorrected observed GLB-values from the sample files
b_p	the mean of the corrected GLB-values from the sample files; it should be close to the population value β
b_z	the mean of the expected GLB-values under the null hypothesis of uncorrelated items
SE(b_p)	the standard error of the mean of the corrected GLB-values
duration	the average time (mm:ss) needed to analyze a single sample file on a basic desk top computer

Table 1. Results of the testing procedure.

test	# files	v	n	β	b_o	b_p	b_z	SE(b_p)	duration
16PF_1	100	10	100	0.6716	0.7559	0.6791	0.3389	0.0075	0:02
16PF_2	100	13	200	0.5581	0.6410	0.5571	0.3099	0.0084	0:06
16PF_3	100	10	500	0.4404	0.4722	0.4373	0.1671	0.0060	0:07
ECR	100	32	100	0.9016	0.9601	0.9052	0.6889	0.0023	1:11
ECR	100	32	200	0.9016	0.9410	0.9044	0.5184	0.0018	1:02
ECR	100	32	500	0.9016	0.9247	0.9072	0.3543	0.0010	1:05
ECR	100	32	1000	0.9016	0.9142	0.9011	0.2545	0.0006	1:09
MSSCQ	50	100	100	0.9675	0.9986	0.9834	0.9725	0.0012	53:36
MSSCQ	50	100	200	0.9675	0.9924	0.9782	0.8406	0.0008	28:20
MSSCQ	50	100	500	0.9675	0.9828	0.9712	0.6138	0.0006	24:44
MSSCQ	50	100	1000	0.9675	0.9772	0.9711	0.4651	0.0003	19:45

The result of these tests strongly suggest that the chosen algorithm reduces the bias in the GLB to a negligible quantity. However, the procedure becomes laborious when the observed GLB is close to unity. It should also be noticed that the expected GLB under the null hypothesis of uncorrelated items (b_z) may become extremely high when the ratio v/n is almost 1.

The assumption of multivariate normality of the items

Above, all scales consisted of ordinal items with a small set of possible scores and their distributions could be estimated from the observed data. If only a covariance matrix is available without information about the distribution of the item scores, one might fall back on the assumption of multivariate normality, but this assumption will frequently be incorrect. In order to get an impression of the seriousness of violations of this assumption, the tests described in the previous section were repeated, now replacing drawings from the actual distributions by drawings from normal distributions. The results are given in Table 2. These results suggest an analysis based on the assumption of multivariate normality will deliver a correct estimator of the GLB, even if the assumption is incorrect.

Table 2. Results using actual distributions and results assuming multinormality.

		Actual Distributions		Normal Distributions	
test	β	b_p	b_z	b_p	b_z
16PF_1	0.6716	0.6791	0.3389	0.6710	0.3397
16PF_2	0.5581	0.5571	0.3099	0.5539	0.3106
16PF_3	0.4404	0.4373	0.1671	0.4353	0.1645
ECR	0.9016	0.9052	0.6889	0.8923	0.6873
ECR	0.9016	0.9044	0.5184	0.9034	0.5206
ECR	0.9016	0.9072	0.3543	0.9066	0.3534
ECR	0.9016	0.9011	0.2545	0.9011	0.2549
MSSCQ	0.9675	0.9834	0.9725	0.9498	0.9576
MSSCQ	0.9675	0.9782	0.8406	0.9723	0.8403
MSSCQ	0.9675	0.9712	0.6138	0.9704	0.6142
MSSCQ	0.9675	0.9711	0.4651	0.9707	0.4659

Table 3. Distributions (proportions) of the 10 items in scale 16PF.

	Items									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Score 1	0.0425	0.0238	0.0350	0.0316	0.0203	0.0185	0.0221	0.0729	0.2560	0.1771
Score 2	0.1163	0.0748	0.0893	0.1087	0.0678	0.0753	0.0699	0.3092	0.4727	0.4053
Score 3	0.1511	0.1787	0.1337	0.1664	0.1767	0.2588	0.1285	0.2486	0.1420	0.2340
Score 4	0.4764	0.4749	0.4771	0.5267	0.4964	0.4732	0.5469	0.2770	0.0997	0.1500
Score 5	0.2137	0.2479	0.2648	0.1666	0.2387	0.1742	0.2326	0.0922	0.0296	0.0336

As an illustration, Table 3 shows the distributions of the items of the population 16PF. The scores are clustered into only 5 categories and the

distribution over these categories is different for the individual items. Nevertheless the estimation of the GLB remains practically unbiased.

Conclusion

It is clear that under the assumptions of the classical test theory and without additional assumptions, the measure known as the Greatest Lower Bound (GLB) is the highest possible lower bound to the reliability of a test. Unfortunately the use of this measure is severely hindered by its bias for small or even moderate samples. It is possible to remove this bias by the given algorithm.

The ideas of this article are implemented in a program called GLBFind, which is available at http://www.ru.nl/socialewetenschappen/rtog/software/statistische/kunst/glbfind/.

References

Bendermacher, N. (2010). Beyond alpha: lower bounds for the reliability of tests. *Journal of Modern Applied Statistical Methods, 9*(1), 95-102.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non- homogeneous items I: Algebraic lower bounds. *Psychometrika, 42*(4), 567-578. doi: 10.1007/bf02295979

personality-testing.info. (n.d.). Raw data from online personality tests. Retrieved from http://personality-testing.info/_rawdata/

Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika, 46*(2), 201-213. doi: 10.1007/bf02293900

Shapiro, A., & ten Berge, J. M. F. (2000). The asymptotic bias of minimum trace factor analysis with applications to the greatest lower bound to reliability. *Psychometrika, 65*(3), 413-425. doi: 10.1007/bf02296154

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 689-721.
doi: 10.22237/jmasm/1493599020

ISSN 1538 − 9472

JMASM45: A Computer Program for Bayesian D-Optimal Binary Repeated Measurements Designs (Matlab)

Haftom T. Abebe
Mekelle University
Mekelle, Ethiopia

Frans E. S. Tan
Maastricht University
Maastricht, The Netherlands

Gerard J. P. van Breukelen
Maastricht University
Maastricht, The Netherlands

Martijn P. F. Berger
Maastricht University
Maastricht, The Netherlands

Planners of longitudinal studies of binary responses in applied sciences have not yet benefitted from optimal designs, which have been shown to improve precision of model parameter estimates, due to absence of a computer program. An interactive computer program for Bayesian optimal binary repeated measurements designs is presented for this purpose.

Keywords: Bayesian optimal designs, logistic mixed effects models, subject-to-measurement cost ratio, relative efficiency, number of time points, autocorrelation

Introduction

Longitudinal study designs are used in different disciplines of science to study the change of a particular outcome variable over time. In smoking prevention studies, for example, pupils in primary and secondary school may be followed up to study the prevalence of smoking as a function of age. The generalized linear mixed model (GLMM) is the most frequently used model for the analysis of longitudinal dichotomous data such as smoking status. Optimal design of longitudinal studies has been shown useful to improve the precision of the model parameter estimates of interest, such as the rate of change, by optimizing the number and timing of repeated measurements. For cross-sectional data, the review of McClelland (1997) provided a good introduction into optimal design for psychologists. Raudenbush and Feng (2001) considered a study with a quantitative outcome in

Haftom T. Abebe is in the Department of Biostatistics, School of Public Health, College of Health Sciences. Email him at: haftom.temesgen@mu.edu.et.

which two groups are followed over time to assess group differences. Optimal design techniques were used to optimize power over feasible designs as a function of duration of a study, frequency of observations, and number of participants. For the GLMM, optimal designs were studied extensively in the literature by Han and Chaloner (2004); Niaparast (2009); Niaparast and Schwabe (2013); and Abebe, Tan, van Breukelen, and Berger (2014a, c), among others.

Unfortunately, optimal designs for nonlinear models depend on the unknown parameter values of interest, that is, on the regression weights that reflect the outcome change over time. Thus, in order to find the optimal design, the model parameter values should be known in advance. However, the parameter values are always unknown as the design is planned to obtained data for estimating them. A common approach to this problem is to use a best guess of the parameter values, which leads to locally-optimal designs, that is, designs which are optimal for a given set of parameter values (see, e.g., Chernoff, 1953). Such designs may not be efficient when the true parameter values differ from those best guesses, that is, the design may not be robust for other parameter values. To overcome this local optimality problem, various methods have been proposed in the literature (see, e.g., Berger & Wong, 2009). The Bayesian approach is one way that has been shown to be useful to take into account the uncertainty of the parameter values (Chaloner & Larntz, 1989; Atkinson, Donev, & Tobias, 2007; Abebe et al., 2014a, b, c; Abebe et al., 2015; among others). The Bayesian design literature is vastly restricted to binary response models. However, no user-friendly software has been developed so far for Bayesian design of longitudinal studies with binary responses.

Due to the absence of a computer program, planners of longitudinal studies in psychology, health sciences, and medicine face the problem of choosing the best number and timing of the repeated measurements. Usually the number and the allocation of the time points at which the measurements are taken are determined by non-statistical criteria. As an example, consider the Dutch smoking prevention study, where smoking and other data were collected from 3735 children in 156 elementary schools by means of a questionnaire at six time points between September 1997 and September 2000: September 1997, February 1998, June 1998, May 1999, February 2000 and September 2000 (Ausems, Mesters, van Breukelen, & De Vries, 2002).

Another example is the attention deficit hyperactivity disorder (ADHD) study (Lahey et al., 1998; Hartung et al., 2002). It was a longitudinal study on 255 children that sought to identify risk and prognostic factors in early childhood for ADHD symptoms, diagnoses, and functional outcomes across childhood,

adolescence, and early adulthood. All participants were followed over seven annual visits after the baseline. The question is whether these designs are efficient, in terms of the number and the timing of the measurements, for estimating the change in smoking and ADHD prevalence over the total follow-up time. This question can be answered by optimal design theory, which is part of the field of statistics.

For the linear random effects model optimal designs were discussed by Tan and Berger (1999) and Tekle, Tan, and Berger (2009), among others. They showed that regardless the underlying polynomial regression model, the number of repeated measures should be chosen as close as possible to the number of regression parameters. Ouwens, Tan, and Berger (2006) and Tekle et al. (2008) extended the work on optimal designs for logistic models with random effects using a maximin approach to handle the local optimality problem, without considering the cost of sampling and measuring. They have kept constant the number of subjects and the number of repeated measures per subject. But in a longitudinal study, costs are associated with the inclusion of patients (subjects) as well as with each repeated measurement.

Further, Bayesian designs are an increasingly popular alternative to maximin design as a method to overcome the local optimality problem. The Bayesian approach takes the uncertainty of the parameter values of the statistical model into account by using a prior distribution on the unknown parameters rather than single-value guesses. This will give more flexibility.

Therefore, a new interactive computer program is presented that computes Bayesian optimal repeated measurements designs for mixed effects logistic models with polynomial time effects under cost constraints, but also allows the user to compute maximin designs. The maximin approach essentially minimizes the largest possible (generalized) variance of the fixed-effect estimators within a user-specified region of the true fixed-effect values, or equivalently, it optimizes among worst possible efficiencies (see, e.g., Tekle et al., 2008; Ouwens et al., 2006).

It computes Bayesian optimal designs for longitudinal studies under cost constraints, thus helping researchers to reduce their study costs. The computer program helps users to identify the optimal number and optimal allocation of time points for a given subject-to-measurement cost ratio. Moreover, it computes the loss in efficiency of equidistant time points compared to the optimal allocation. It produces a plot of optimal allocations of time points under different values of autocorrelation. A separate manual is presented in the appendix and describes the

capabilities of the software, which runs in a Matlab environment (MathWorks, 2010).

The logistic mixed effects model with polynomial time effects is described, and the optimality criterion and the relative efficiency as a measure for the comparison of designs. Thereafter, the smoking prevention study by Ausems et al. (2002) is used to illustrate the application of the program and to discuss the various decisions that the user has to make when determining the most efficient design. The manual can be considered as part of the paper, but can be consulted independently from it. Finally, conclusions and recommendations are provided. The paper ends with a summary and discussion.

The Logistic Mixed Effects Model

Let the $q \times 1$ vector $\mathbf{y}_i = (y_{i1},\ldots, y_{iq})'$ be binary responses y_{ij} of subject i at q time points, $i = 1, 2,\ldots, N$ and $j = 1,\ldots, q$. It is assumed that all subjects have measurements at the same time points, and that, conditional on the subject-specific random effect vector $\mathbf{b}_i$, the binary responses y_{ij} of $\mathbf{y}_i$ are assumed to be Bernoulli distributed with probability of success $\mathrm{p}(y_{ij} = 1 \,|\, \mathbf{b}_i)$. These probabilities are related to the fixed and random effects via the logit link function. The corresponding logistic mixed effects model is given by:

$$\operatorname{logit}\left(\mathrm{p}\left(y_{ij}=1|\,\mathbf{b}_i\right)\right)=\log\left(\frac{\mathrm{p}\left(y_{ij}=1|\,\mathbf{b}_i\right)}{1-\mathrm{p}\left(y_{ij}=1|\,\mathbf{b}_i\right)}\right)=\mathbf{x}'_j\boldsymbol{\beta}+\mathbf{z}'_j\mathbf{b}_i \qquad (1)$$

where the $p \times 1$ vector $\mathbf{x}_j$ is the design vector of the explanatory variables at the j^{th} measurement for subject i, $\boldsymbol{\beta}$ is the corresponding $p \times 1$ vector of fixed polynomial time effects, and $\mathbf{z}_j$ is the $r \times 1$ design vector for the random effects that is usually a subset of vector $\mathbf{x}_j$. The vector $\mathbf{b}_i$ is the corresponding $r \times 1$ vector of random effects, which is assumed to have a multivariate normal distribution with mean zero and covariance matrix $\mathbf{D}$.

For example, if a quadratic ($p = 3$) time effect is assumed, the design vector is $\mathbf{x}'_j = \begin{pmatrix}1 & t_j & t_j^2\end{pmatrix}$ and $\boldsymbol{\beta} = (\beta_0\ \beta_1\ \beta_2)'$, where t_j is the time point of the j^{th} measurement, $j = 1,\ldots, 6$, and β_0, β_1, and β_2 are the fixed effects. Suppose that a random intercept and random linear slope are assumed. Then the design vector is $\mathbf{z}'_j = (1\ t_j)$ and $\mathbf{b}_i = (b_{0i}\ b_{1i})'$, where b_{0i} and b_{1i} are the corresponding random (subject-specific) deviations from these fixed effects, $i = 1,\ldots, 3735$. Then,

according to model (1), the log-odds of a positive response (smoking) for subject i at time t_j is given by:

$$\text{logit}\left(\text{p}\left(y_{ij}=1|\mathbf{b}_i\right)\right)=\left(\beta_0+b_{0i}\right)+\left(\beta_1+b_{1i}\right)t_j+\beta_2 t_j^2 \qquad (2)$$

To prevent misunderstanding about the flexibility of this model, note that it can handle U-shaped as well as monotonic trends over time. For the average subject (i.e. if the random effects are zero), the derivative of (2) with respect to time t is 0 if

$$t=\frac{-\beta_1}{2*\beta_2}$$

The time variable is bounded by the follow-up period of the longitudinal study, and so equation (2) reaches its maximum or minimum inside or outside the time interval, depending on the values of β_1 and β_2. So model (2) can handle monotonic as well as non-monotonic trends.

For example, in the Dutch smoking prevention study, a quadratic ($p = 3$) time effect will be needed if smoking prevalence on the logodds scale increases nonlinearly over time. For the sequel, it is important to note that in this paper and software, the time interval is scaled as $t \in [-1, +1]$. This can be translated into any suitable time scale by linear transformation, and vice versa. For instance, the time scale of the smoking prevention study, with its baseline of September 1997 as the origin, its last measurement in September 2000, and a month as the unit of measurement, is obtained by the transformation $t^* = 18(t + 1)$. Likewise, our present time scale is obtained as $t = (t^* - 18)/18$. The repeated measurements of smoking were made at time points $t^* = 0$, 5, 9, 20, 29, and 36 months, which in terms of the present time scale gives as time points t = -1.00, -0.72, -0.50, 0.11, 0.61, and 1.00, respectively.

Due to the random effects in model (1) and (2), the log-likelihood cannot be written down in closed form. Hence, either numerical methods or approximations to the log-likelihood must be used. Numerical methods require large computational resources and more importantly they require full knowledge of the data (Moerbeek, Van Breukelen, & Berger, 2003; Han & Chaloner, 2004), making them computationally inconvenient for optimal design procedures. To overcome this problem, approximation methods are employed. There is a large statistical literature on various approximation methods, but here, for the purpose

of obtaining optimal designs, we will focus on the two most frequently used ones, which are implemented in commercially available software packages: first order penalized quasi-likelihood (PQL1) and an extended version of generalized estimating equations (GEE).

First Order Penalized Quasi-Likelihood

The PQL1 variances and covariances of the fixed parameter estimates are calculated using the first-order Taylor expansion around the fixed and random effects. An advantage is that the method performs well in terms of point estimates since it produces the smallest mean squared error and the bias of the estimators decreases as the sample size increases (Breslow & Clayton, 1993; Moerbeek et al., 2003; Jang & Lim, 2009). A disadvantage is that design optimization based on PQL1 is very time consuming. This is due to the fact that the covariance matrix of the binary responses, which must be inverted at each iteration of the optimization process, is very large because it depends on the random effects (which in the design stage are sampled from a multinormal distribution). The variance-covariance matrix of the estimator $\hat{\boldsymbol{\beta}}$ of the parameter $\boldsymbol{\beta}$ for the logistic mixed effects models (1) is approximated in PQL1 by:

$$\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right) \approx \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \tag{3}$$

where $\mathbf{X}$ is the $Nq \times p$ design matrix formed by stacking $\{\mathbf{x}'_j\}$ for N subjects and q time points, and $\mathbf{V}$ is the $Nq \times Nq$ block-diagonal matrix with N blocks of $q \times q$ variance-covariance matrices given by:

$$\mathbf{v}_i \approx \mathbf{w}_i^{-1/2}\mathbf{R}\left(\rho\right)\mathbf{w}_i^{-1/2} + \mathbf{ZDZ}' \tag{4}$$

The $q \times q$ matrix $\mathbf{R}(\rho)$ is the residual correlation matrix, $\mathbf{Z}$ is the $q \times r$ design matrix with rows $\mathbf{z}'_j$, $j = 1,\ldots, q$, the $r \times r$ matrix $\mathbf{D}$ is the variance-covariance matrix of the random effects, and $\mathbf{w}_i^{-1}$ is the diagonal matrix of the conditional variances of the transformed responses given the random effects $\mathbf{b}_i$, which is equal to the inverse of the diagonal matrix of the conditional variances of the untransformed responses given the random effects $\mathbf{b}_i$ (See for detail Moerbeek et al., 2001; Molenberghs & Verbeke, 2005, p. 270). Note that, under conditional independence, $\mathbf{R}(\rho)$ is an identity matrix and equation (4) becomes

$$\mathbf{v}_i \approx \mathbf{w}_i^{-1} + \mathbf{ZDZ}' \tag{5}$$

The diagonal matrix of the conditional variances of the untransformed responses given the random effects $\mathbf{b}_i$, is given by:

$$\mathbf{w}_i = \text{diag}\left(w_{i1}^{\mathbf{b}_i}, \ldots, w_{iq}^{\mathbf{b}_i}\right) \tag{6}$$

where $w_{ij}^{\mathbf{b}_i} = \text{var}\left(y_{ij} \mid \mathbf{b}_i\right)$, for $i = 1,\ldots, N$, $j = 1,\ldots, q$. Since the random effects are unknown in the design stage, we will generate $\mathbf{b}_i$ from a multivariate normal distribution with mean zero and variance-covariance $\mathbf{D}$.

Extension of Generalized Estimating Equations

The extended GEE is an alternative method which is not likelihood-based. It has been extended by Zeger, Liang, and Albert (1988) and Molenberghs and Verbeke (2005) to include autocorrelations of the errors in the standard formulation of GLMM. The covariance matrix of the binary responses is expressed conditional on the random effects being zero, which makes the calculations much faster. The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ for the logistic mixed effects models (1) with autocorrelation, based on the extension of the GEE approach, is approximated by:

$$\text{var}\left(\hat{\boldsymbol{\beta}}\right) \approx \left(\sum_{i=1}^{N} \frac{\partial^2 \mathbf{P}_i'}{\partial \boldsymbol{\beta}} \mathbf{u}_i^{-1} \frac{\partial^2 \mathbf{P}_i}{\partial \boldsymbol{\beta}'}\right)^{-1} \tag{7}$$

where $\hat{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}$ for model (1), $\mathbf{P}_i = (\text{p}(y_{i1} \mid \mathbf{b}_i),\ldots, \text{p}(y_{iq} \mid \mathbf{b}_i))'$ and the working variance-covariance matrix of the responses is given by:

$$\mathbf{u}_i \approx \mathbf{w}_i^{1/2} \mathbf{R}\left(\rho\right) \mathbf{w}_i^{1/2} + \mathbf{w}_i \mathbf{ZDZ}' \mathbf{w}_i \tag{8}$$

When there are no residual correlations in $\mathbf{R}(\rho)$, a conditional independence model or purely random effects model results and equation (8) reduces to

$$\mathbf{u}_i \approx \mathbf{w}_i + \mathbf{w}_i \mathbf{ZDZ}' \mathbf{w}_i \tag{9}$$

where $\mathbf{w}_i$ is the diagonal matrix of the conditional variances of untransformed responses given the random effects $\mathbf{b}_i = 0$, which is given by:

$$\mathbf{w}_i = \text{diag}\left(w_{i1}^{\mathbf{b}_i=0}, \ldots, w_{i1}^{\mathbf{b}_i=0}\right) \tag{10}$$

where $w_{ij}^{\mathbf{b}_i=0} = \text{var}\left(y_{ij} \mid \mathbf{b}_i = 0\right)$ for $i = 1, \ldots, N$, $j = 1, \ldots, q$ (Molenberghs & Verbeke, 2005, p. 443).

Time-structured data are naturally correlated (Berger, 1986). In this paper, a first order auto regressive (AR1) is considered, i.e., $\rho^{|t_j - t_l|}$, where $j, l = 1, \ldots, q$, and so ρ is the autocorrelation coefficient between two responses at a time distance of one, that is, $\rho = \text{Corr}(y_{ij}, y_{il})$ for which $|t_j - t_l| = 1$. This autocorrelation structure implies that repeated measurements closer in time are more highly correlated and that the correlation decreases as the distance between the time points increases.

Bayesian D-Optimal Design and Relative Efficiency

To introduce the notation for the optimality criterion, suppose that the study to be designed will have q ordered time points $t_1, t_2, \ldots, t_q$ at which measurements are taken for all N subjects. The design space Ξ then contains all designs of the form

$$\Xi = \left\{ \begin{pmatrix} t_1 & t_2 & \ldots & t_q \\ w_1 & w_2 & \ldots & w_q \end{pmatrix} : t_j \in [a, b], t_1 < t_2 < \ldots < t_q \right\} \tag{11}$$

with weight w_i indicating per time point what proportion of all observations is obtained at that point (see also, e.g., Bunke & Bunke, 1986, p. 506) and $q \geq p$ to make these fixed effects identifiable with p being the number of fixed parameters of the model. Although in general the weights (w_i) at the different time points can be different, in this paper we make the restriction of all weights equal to 1 ($w_1 = w_2 = \ldots = w_q = 1$) at all q ordered time points, i.e., measurements are taken on all N subjects at each time point, because we consider longitudinal designs and so all q repeated measurements are obtained from the same individuals. The time interval $[a, b]$ is assumed to be fixed by substantive constraints within the field of application, for example, the total follow-up time in the cohort study of smoking prevention is $b - a = 3$ years, or 36 months. A design ξ_q is an element of the design space Ξ if it has q time points within the time interval $[a, b]$.

Optimal designs are usually selected by minimizing a real-valued function of the variance-covariance matrix of the parameter estimators, here of the estimators of the three regression weights in (2), which is known as optimality criterion (see, e.g., Silvey, 1980). In this way the precision of the estimators and the power of their significance tests are maximized. Various optimality criteria have been proposed in the literature, such as the D-, A-, or G-optimality criteria. In this paper, we will focus on the best-known and most popular optimality criterion, i.e., the D-optimality criterion. This optimality criterion has two nice properties: 1. It minimizes the volume of the asymptotic confidence ellipsoid for the parameters, for instance for the fixed effects in model (2), thus giving the multivariate generalization of the familiar confidence interval for a single parameter; and 2. It does not depend on the coding used for the endpoints of the chosen time interval [a, b], for instance, on whether we code the time predictor in equation (2) as running from 0 to 1, or from -1 to +1, or use the original time scale in days or months. This means that if the coding for the time interval is transformed linearly, a D-optimal design for the new time interval is obtained by applying the same linear transformation to the D-optimal design for the old interval (see Ouwens et al., 2006).

For example, in the smoking study, the measurements were taken between September 1997 and September 2000 (a period of three years), and by linearly transforming the measured time points into the interval [-1, +1], the actual design of the smoking study ξ_6 becomes (-1 -0.72 -0.50 0.11 0.61 1). Likewise, if e.g. the D-optimal allocation of the time points for the smoking study is -1, -0.5, 0, 0.5, and 1 on the time interval [-1, +1], then it is after 0, 9, 18, 27, and 36 months respectively on the original time scale of [0, 36] months.

The D-optimal design ξ_q^* is the design among all possible designs ξ_q with q time points for which the determinant of the variance-covariance matrix of parameter estimators, for instance, the covariance matrix of $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ in model (2), is minimized (Berger & Wong, 2009). It should be noted that, for some studies, other criteria could be more obvious. Using the D-optimality criterion, all fixed-model parameters are considered to be equally important. If, for example, only some of the model parameters are of interest and others are considered to be nuisance, then a D_A-criterion will be more relevant, indicating that only a subset or m linear combinations of the p regression parameters ($p \geq m$) are of interest and specified by an $m \times p$ design matrix $\mathbf{A}$ (see, e.g., Tan, 2011). Nevertheless, the concentration here will be on this D-optimality criterion, because it is expected all fixed effects in model (1) will be of interest.

The variance-covariance matrix of the fixed-effects estimators $\hat{\boldsymbol{\beta}}$ depends on the unknown parameter vector $\boldsymbol{\beta}$ (see Abebe et al., 2014a, b, c; Abebe et al., 2015), which makes design optimization dependent on the very same parameters that have to be estimated with the study to be designed, thus creating a vicious circle. The Bayesian approach resolves this dependency problem by taking the expectation of a function of the variance-covariance matrix over a prior distribution for the unknown parameter vector $\boldsymbol{\beta}$. Thus, the Bayesian D-optimality criterion is defined as follows:

$$\begin{aligned}\phi_D(\xi \mid \pi) &= E_{\boldsymbol{\beta}} \log\left|\left(\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}\right| \\ &= \int_{\boldsymbol{\beta}} \log\left|\left(\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}\right| \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta}\end{aligned} \tag{12}$$

where $\pi(\boldsymbol{\beta})$ is the prior distribution for $\boldsymbol{\beta}$ and $\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right)$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ for the logistic mixed effects models based on approximation methods (see Abebe et al., 2014a, b, c; Abebe et al., 2015 for details). In fact, the design criterion (12) follows from maximizing the expected Kullback-Leibner (KL)-distance between the prior and posterior distributions, measuring how much information can be gained when moving from prior to posterior. When the normal approximation is used for the posterior distribution, then a design that maximizes the KL-distance is equivalent to maximizing expression (12) and is called Bayes D-optimal. It should be mentioned that expression (12) does not represent the full Bayesian design criterion, but only approximately by ignoring the additional effect of the prior information about the fixed effects. However, for large sample sizes, the contribution of the prior information to the posterior variance is usually negligible (for further details, see Chaloner & Verdinelli, 1995; Sebastiani & Settimi, 1998). Note that maximization of (12) comes down to minimization of the expected log determinant of the covariance matrix, where the expectation is taken over the prior (Atkinson et al., 2007).

The precision of estimating the fixed-effects parameters $\boldsymbol{\beta}$ increases by taking more measurements and sampling more subjects (Moerbeek et al., 2001). However, the addition of subjects and of measurements per subject will increase the costs of the study and these are usually limited by budget constraints. Therefore, it is reasonable to take into account the costs of a longitudinal study when designs are compared with each other. There are two main components of

these costs. These are the costs for recruitment of subjects and the costs of the measurements once a subject has been recruited. Let the cost of recruiting a subject be C_1 and the cost of one measurement per subject be C_2. Then the total cost of a longitudinal study with q time points and N subjects, excluding overhead cost, is given by the linear cost function:

$$\begin{aligned} C &= C_1 N + C_2 N q \\ &= N C_2 \left(k + q\right) \end{aligned} \tag{13}$$

where $k = C_1/C_2$ is the ratio of the cost of adding a new subject to the cost of an additional measurement per subject.

To compare different designs, we will use their relative efficiencies while fixing the total costs C. This means that the designs can differ in terms of the number of subjects N and the number and timing of the measurements q. First, we compute the Bayesian D-optimal designs using fixed N and then we correct for costs and different q and N as follows: Let $\phi_{\mathrm{D}}\left(\xi_q^* \mid \pi\right)$ denote the value of design criterion (12) for the optimal design ξ_q^* with q time points, given the prior distribution π for the fixed effects. Then the relative efficiency (RE) of an arbitrary design ξ_s with s time points relative to the optimal design ξ_q^* is defined as:

$$\mathrm{RE}\left(\xi_s; \xi_q^*\left(\pi\right) \mid \pi\right) = \frac{k+q}{k+s}\left[\exp\left\{\frac{\phi_{\mathrm{D}}\left(\xi_s \mid \pi\right) - \phi_{\mathrm{D}}\left(\xi_q^* \mid \pi\right)}{p}\right\}\right] \tag{14}$$

where π is the prior distribution for the fixed effects and p is the number of fixed effects, that is, $p = 3$ for model (2). If the value of this relative efficiency is close to unity, then the design ξ_s is about equally efficient as the optimal design ξ_q^* for a given prior π. The inverse of this relative efficiency is the number of times that a design ξ_s must be replicated to have the same efficiency as the optimal design ξ_q^*. Note that the term between squared brackets on the right side of equation (14), so without the $(k + q)/(k + s)$ term, is the RE under the assumption of an equal number of subjects N for both designs, which then differ only in the number and timing of the repeated measures. This fixed N-situation, i.e. $N_s = N_q$, underlies the RE formula as given by Chaloner and Larntz (1989). However, if we keep the

total budget *C* instead of *N* the same for all designs, then it follows from equation (13) that we can have

$$\frac{N_s}{N_q} = \frac{k+q}{k+s}$$

as many subjects in design ξ_s as in design ξ_q^*. Since $\text{var}\left(\hat{\boldsymbol{\beta}}\right)$ is inversely proportional to the sample size *N*, it then follows from equations (12) and (13) that the RE of both designs obeys equation (14). See the Appendix for details on the derivation of RE in (14).

Method of Optimization

The Bayesian D-optimal designs for the logistic mixed effect model are found by our computer program numerically by maximization of the criterion value (12) among all candidate designs for a given prior distribution of the parameters. Details of this will be given in the next sub-sections.

Sampling Parameter Values from Priors to Compute the Criterion

To construct Bayesian designs for continuous prior distributions, all candidate designs must be evaluated in terms of their criterion values as defined by (12). However, evaluation of the integration over the prior distribution is very complicated and cannot easily be done analytically. A numerical approximation of the integral is necessary. Numerical approximations can be done by sampling parameter values from the prior distribution and then by replacing the integral in (12) with a summation over the sample (Atkinson et al., 2007; Chaloner & Verdinelli, 1995). Estimating (12) using the traditional sampling (pseudo Monte Carlo) method requires very large samples from the prior to reduce the sample-to-sample variability to the point where different samples do not lead to different design choices. Thus, this approach is costly in terms of computing time. In our computer program, we will use an Adaptive Rejection Metropolis Sampling (ARMS) algorithm (Gilks & Wild, 1992; Gilks, Best, & Tan, 1995), which is a more efficient sampling algorithm that requires a smaller sample to obtain a good approximation of the design criterion (12). ARMS is a generalization of the method of adaptive rejection sampling (ARS) (Gilks, 1992), which was itself a development of the original method proposed by Gilks and Wild (1992). The

ARMS generalization includes a Metropolis step to accommodate non-concavity in the log density. ARMS is a Markov chain Monte Carlo (MCMC) scheme for generating samples from high dimensional target distributions and widely used within Gibbs sampling, where automatic and fast samplers are often needed to draw. It can deal with (intrinsic) non-linear functions as often used in, for instance, pharmacokinetics. For the present log-linear model, the ARMS works very well and is much faster than the Gibbs sampling method.

Optimization Algorithm for Finding an Optimal Design

To find candidate designs and in particular the optimal design, the program uses the FMINCON function of MATLAB version 7.10.0499 (R2010a). This function performs constrained non-linear optimization and requires an initial design ξ_0. Without loss of generality, the time interval was coded as [-1, +1], and equally-spaced time points were used as initial designs. There is no need to start with non-equally spaced time points because our experience is that Bayesian optimal designs for our model do not depend on the spacing of the initial design. According to Firth and Hinde (1997), the Bayesian criterion may only lead to different optimal designs for different starting values when very dispersed prior distributions are considered. In fact, the Bayesian D-optimal designs as obtained with our program can deviate a lot from equidistance, thus showing that equidistance as initial design does not constrain the final design (see, e.g., Abebe et al., 2015).

The following global search algorithm is used to find the Bayesian D-optimal designs for a given multivariate normal prior distribution of the parameters:

1. Take samples from the prior distribution of the parameters using ARMS.
2. Compute the Bayesian D-optimal allocation of q time points, using $q = p$ equidistant time points as initial design, where p is the number of fixed parameters of the model. Note that the final optimal allocation does not need to be equally spaced (see, e.g., Abebe et al., 2014a).
3. Increase the number of time points q by one and perform step 2 again to find the Bayesian optimal design (allocation) for the new value of q. Repeat step 2 and 3 until the maximum number of time points q (user specified) is reached.

4. Thereafter, select the optimal number of time points q for the Bayesian D-optimal design by computing the relative efficiencies of designs with different numbers of time points against each other for a user-specified subject-to-measurement cost ratio. Do this for each cost ratio considered to obtain one optimal design per cost ratio for a chosen prior distribution.

An Example: The Dutch Smoking Prevention Study

As an illustration of the various decisions that the user has to make when determining the most efficient design, consider the Dutch smoking prevention study as described in the introduction section. A logistic mixed-effects model with quadratic time effect was found to give an adequate fit to the repeated measures of smoking status (0 = no, 1 = yes). Therefore, this model was adopted to illustrate the application of the BODMixed_Logistic program in guiding researchers for a similar future study. After starting the BODMixed_Logistic program, all the steps will be reviewed that are necessary to obtain the optimal design, starting with the specification of the model and the various input values. See the program manual for a description of the graphical user interface offered per step.

Choice of the Model

The first step is to choose the statistical model; the optimal design depends on the underlying statistical model and is different for a quadratic model than for a linear one. For the fixed model part, we choose a quadratic growth function, both in view of its fit to the smoking data and because it is more flexible than a linear one and can handle monotonic trends as well as U-shaped trends due to the finite time interval. For the random model part, we assume a random intercept as well as a random linear slope. This can be specified in the program by choosing nonzero variances for the intercept and linear slope and zero variance for the quadratic slope, with or without slope-intercept covariance.

To the program user it may be reassuring to know that Abebe et al. (2014c) found that the Bayesian D-optimal designs are hardly affected by the choice of a covariance structure for the random effects, at least in case of a non-zero autocorrelation and the presence of a random intercept or random slope. Further, the autocorrelation between the repeated measures must be specified. Fortunately, the maximum loss in efficiency incurred by misspecification of the autocorrelation appears to be less than 5% (Abebe et al., 2014c), excepting the

case of a zero autocorrelation which gives very different allocations of time points than nonzero values. For illustration purpose, we will assume the default value of 0.1 for the autocorrelation, remembering that this is the correlation between two measurements with a time interval of 1 on the time scale [-1, +1]. Of course, the program user is free to try out different covariance structures and autocorrelations to check the dependence of the optimal design on these values for his/her specific study.

Approximation Method

Next, the user has to choose between the two approximations of the likelihood that are implemented in the program: PQL1 and extended GEE. If computation time is not an issue, then we would recommend using the PQL1 approximation. The extended GEE, however, is computationally much faster and often produces similar Bayesian D-optimal designs as the PQL1 approximation (Abebe et al., 2014c). In this example, we choose the extended GEE.

Choice of Optimality Criterion

At this stage, the model and the necessary parameter values have been specified. The program offers three different optimization criteria.

a. The option 'Bayesian D-optimal' maximizes the criterion in equation (12), thus minimizing the generalized variance of the fixed effects estimators, for a user specified prior distribution of those fixed effects. Abebe et al. (2014b, c) showed that it is best to choose a prior distribution with a large variance (uninformative prior) to express the degree of uncertainty about the 'true' parameter values. The prior means then have little impact on the optimal design, provided that the autocorrelation is not too close to zero ($\rho > 0.001$).
b. The option 'locally D-optimal' criterion can be chosen if the user wants to check the optimal design for specific values of the fixed effects regression parameters. Note that this comes down to assuming a prior with zero variance. This option is in general not recommended, because it will often lead to a sub-optimal design.
c. The option 'Maximin D-optimal design' essentially minimizes (among all possible designs) the largest possible (generalized) variance of the fixed-effect estimators within a user-specified region of the true fixed-effect values, or equivalently, it maximizes the minimum efficiency within this

> region (see, e.g., Tekle et al., 2008; Ouwens et al., 2006). Using this criterion, the user remains on the safe side, and will furthermore obtain a design that is optimal for at least one combination of likely parameter values. A disadvantage of this criterion is that the maximin design is often optimal for some points on the boundary of the region ("parameter space") for the true fixed effects, and these boundary points are less likely than values within the region (Atkinson et al., 2007, p. 258).

For this illustration, Bayesian D-optimal design is selected with, as input prior distribution for the fixed effects, an independent normal with prior means μ = [1, 2, 3] and a prior variance $\sigma^2 = 5$ for both fixed effects. Abebe et al. (2014c) showed that the Bayesian D-optimal designs with such large prior variance are hardly affected by the choice of prior means, provided that the autocorrelation is not too close to zero ($\rho > 0.001$).

Optimal designs can be determined now in either of two ways: By fixing the number of time points q and finding the optimal q allocations, or by finding the optimal number and allocation of time points for a given subject-to-measurement cost ratio k.

Computing the Optimal Allocation for a Given Number of Time Points *q*

For this illustration, we use $q = 6$ time points as the design in the smoking example had 6 repeated measurements. The resulting optimal time points are, according to Figure 1 (see the 4th design in it), [-1, -0.6080, -0.2063, 0.1875, 0.5465, 1]. Translated into the scale of the smoking study period in months, that is, into the time interval [September 1997, September 2000], this gives as optimal design points September 1997, April 1998, November 1998, June 1999, January 2000, and September 2000. To compare, the actual time points were September 1997, February 1998, June 1998, May 1999, February 2000, and September 2000. In this example we fixed the number of time points, but it may be of interest to find the optimal number of time points for a given subject-to-measurement cost ratio, which will now be discussed.

Finding the Optimal Design for a Given Subject-to-Measurement Cost Ratio *k*

As mentioned previously, the user can choose between fixing the number of time points q and fixing the subject-to-measurement cost ratio. The second option will

now be illustrated assuming a cost ratio $k = 1$, that is, equal costs for recruiting a subject and for a single measurement on a single subject. A maximum of seven time points were chosen, which covers the number of time points in most longitudinal studies. The minimum is three because the model has $p = 3$ fixed effects and is thus not identifiable with less than three time points.

The results are given in Figure 1, showing the Bayesian optimal designs for each of the number of time points $q = 3, 4,\ldots, 7$, and the relative efficiency of each Bayesian optimal design compared to the Bayesian optimal design with $q = 7$ time points for the chosen cost ratio, here $k = 1$. The optimal number of repeated measures q for that cost ratio is $q = 4$, giving a relative efficiency of 1.2324 compared to $q = 7$. Further, the relative efficiency of an equidistant design with $q = 4$ time points compared to the optimal design with $q = 4$ is 0.9770, and so equidistance is highly efficient here, although it is not optimal. Finally, to show the effect of the chosen cost ratio on the optimal design, Figure 2 gives a plot of the relative efficiencies of the Bayesian optimal designs with different numbers of time points compared to the optimal design with the maximum number of time points, for each of several cost ratios k. Clearly, the optimal number of time points increases as the subject-to-measurement cost ratio becomes large. The practical implication of this is that, if the user is uncertain about the cost ratio, he or she should try several cost ratios within the plausible range.

The efficiencies of the actual design of the smoking design relative to the Bayesian optimal design increase with an increasing cost ratio k, and the relative efficiency is large for cost ratios $k \geq 2$. For small cost ratios k, the loss in efficiency for the actual design relative to the Bayesian design with 4 time points is at most 25%, which can be compensated by sampling about 33% more children. For large cost ratios ($k \geq 10$), the loss in efficiency for the actual design is at most about 4%, which can be compensated by sampling about 4% more children.

Plotting the Bayesian Optimal Design for Different Values of the Autocorrelation

In the example it was assumed there is a single value 0.1 for the autocorrelation. However, the autocorrelation is rarely known in the design stage. The program therefore offers as a last option a plot of the effect of the autocorrelation value on the Bayesian D-optimal design for a user specified number of time points q and range of autocorrelation. Figure 3 shows such a plot for $q = 6$ time points (horizontal axis) against the autocorrelation (vertical axis) within the range from 0.001 to 0.90 for the random intercept logistic model with

quadratic time effects. From this plot we see that the Bayesian D-optimal allocation for $q = 6$ is fairly independent of the size of the autocorrelation, at least within the chosen range from 0.001 to 0.9. As mentioned before, a zero autocorrelation usually gives quite different optimal allocations which are far from equidistant.

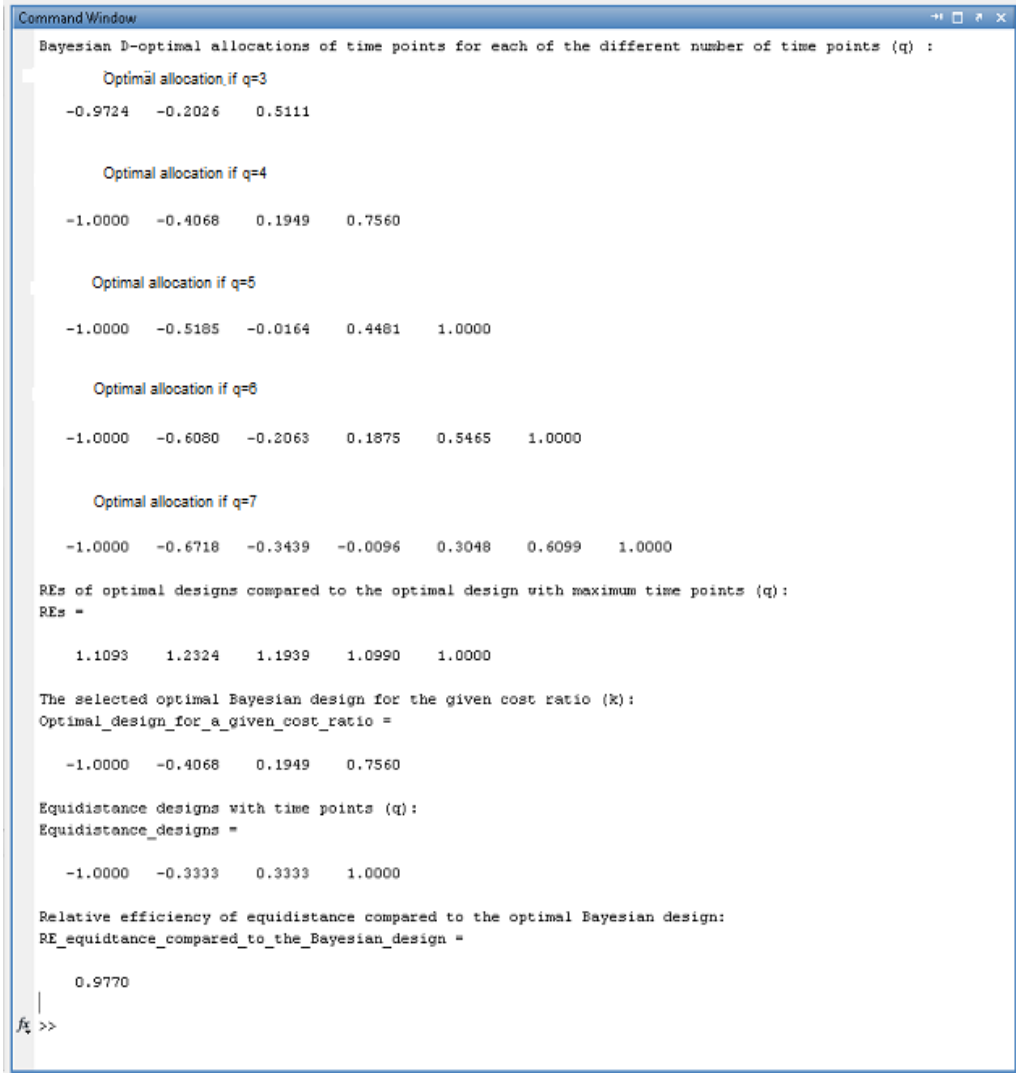


Figure 1. Bayesian optimal allocations of time points for cost ratio $k = 1$ with a maximum number of time points $q = 7$ for the logistic mixed model with quadratic time effects, assuming a random intercept and random linear slope logistic model with quadratic time effects, and autocorrelation 0.1

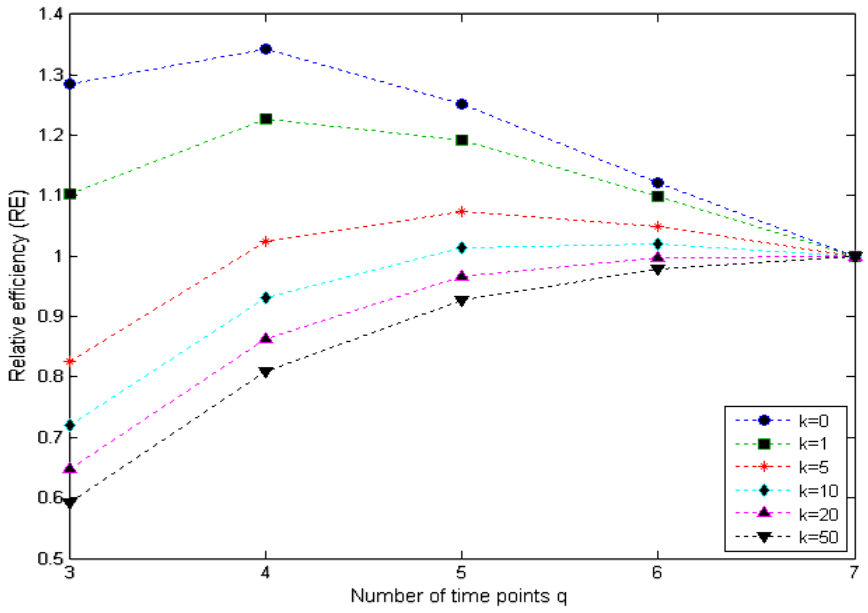


Figure 2. Relative efficiency of Bayesian optimal designs compared to the Bayesian optimal design

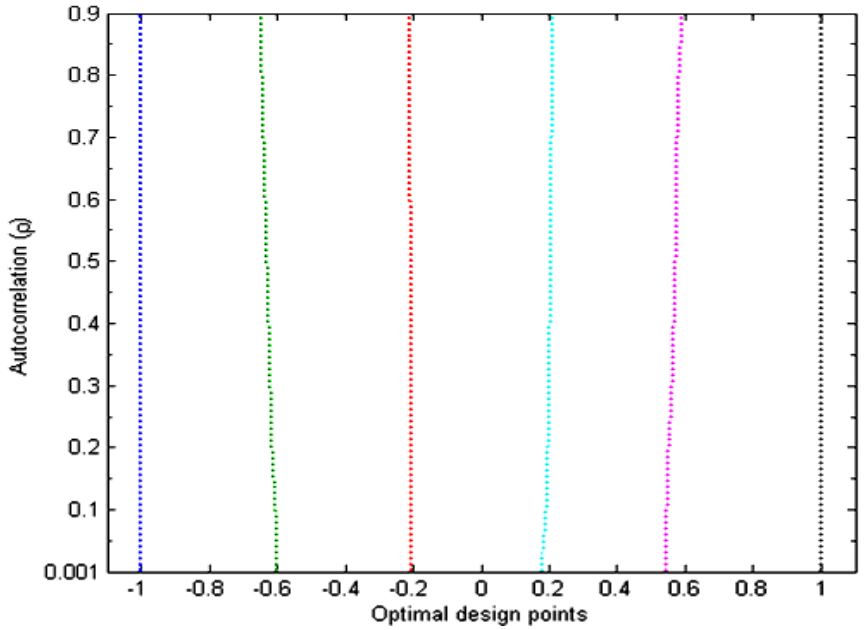


Figure 3. Bayesian D-optimal allocation of $q = 6$ time points as a function of the autocorrelation, for the logistic mixed model with quadratic time effects

Table 1. Optimal time points for a quadratic model, fixed effects, number of time points $q = 6$

Prior variance σ^2	Prior mean ($\beta_1, \beta_2, \beta_3$)	Autocorrelation (ρ) 0.0	0.01	0.9
0.5	[0, 0, 0]	(-1, -1, 0, 0, 1, 1)	(-1,-0.60,-0.19, 0.19,0.60,1)	(-1,-0.63,-0.22,0.22,0.64,1)
	[1, 2, 3]	(-1,-1, -0.27, -0.27, 0.47,0.47)	(-1,-0.66,-0.30,0.04,0.40,0.71)	(-1,-0.69, -0.25,0.21,0.53,1)
5	[0, 0, 0]	(-1, -0.36, 0.06, 0.06, 0.56, 1)	(-1, -0.60, -0.20, 0.20, 0.60, 1)	(-1,-0.67,-0.23,0.25,0.67, 1)
	[1, 2, 3]	(-1,-0.60,-0.24,0.13,0.46,0.91)	(-1, -0.60, -0.21, 0.18, 0.54, 1)	(-1,-0.65,-0.21,0.21, 0.60,1)

Table 2. Optimal time points for a quadratic model, random intercept, intercept variance $\tau_0^2 = 1$, number of time points $q = 6$

Prior variance σ^2	Prior mean ($\beta_1, \beta_2, \beta_3$)	Autocorrelation (ρ) 0.0	0.01
0.5	[0, 0, 0]	(-1, -1, 0, 0, 1, 1)	(-1, -0.58, -0.18, 0.18, 0.58, 1)
	[1, 2, 3]	(-1, -1, -0.26, -0.26, 0.51, 0.51)	(-1, -0.66, -0.29, 0.06, 0.41, 0.73)
5	[0, 0, 0]	(-1, -1, -0.28, 0.29, 1, 1)	(-1, -0.60, -0.20, 0.20, 0.60, 1)
	[1, 2, 3]	(-1, -0.62, -0.22, 0.19, 0.50, 0.98)	(-1, -0.60, -0.21, 0.18, 0.55, 1)

Table 3. Optimal time points for a quadratic model, random intercept/slope, random intercept variance $\tau_0^2 = 1$, random slope variance $\tau_1^2 = 1$, number of time points $q = 6$

Prior variance σ^2	Prior mean ($\beta_1, \beta_2, \beta_3$)	Autocorrelation (ρ) 0.0	0.01
0.5	[0, 0, 0]	(-1, -1, 0, 0, 1, 1)	(-1, -0.56, -0.17, 0.17, 0.57, 1)
	[1, 2, 3]	(-1, -1, -0.24, -0.24, 0.51, 0.51)	(-1, -0.64, -0.28, 0.07, 0.42, 0.73)
5	[0, 0, 0]	(-1, -0.66, -0.17, 0.17, 0.63, 1)	(-1, -0.60, -0.20, 0.20, 0.60, 1)
	[1, 2, 3]	(-1, -0.59, -0.20, 0.18, 0.50, 0.97)	(-1, -0.60, -0.21, 0.18, 0.54, 1)

Summarizing the example of Bayesian optimal design with the BODMixed_Logistic program, it can be concluded that when the subject-to-measurement cost ratio k is less than 5, i.e. the cost of an additional subject does not exceed five times the cost of an additional observation on a single subject, then the optimal number of repeated measurements is four time points. Further, the optimal allocation is not equidistant, but equidistance is highly efficient.

Using the suggested Bayesian D-optimal design, the relative efficiency of the optimal number of repeated measures q for the given cost ratio $k = 1$, which is equal to $q = 4$, relative to the $q = 6$ (which is the number of time points in the smoking study) is equal to about RE = 1.2324/1.099 = 1.1213 (see Figure 2). This means that about 10% less budget is needed for the optimal design to reach the same efficiency as compared to the actual design of the smoking prevention study of Ausem et al. (2004), which had six time points.

Finally, to demonstrate the effect of the covariance structure D, prior means and variances, as well as of autocorrelation on the Bayesian D-optimal design, we will show some additional results for a quadratic model with fixed effects, random intercept, random intercept/slope, and for various priors and autocorrelations. We fixed the number of time points to $q = 6$ and used the extended GEE method for these results which are summarized in Tables 1 to 3, which gives the optimal time points for varying parameter values.

Shown in Table 1 are optimal allocations of time points for a quadratic model with fixed effects only, Table 2 for the random intercept model with intercept variance equal to $\tau_0^2 = 1$, and Table 3 for the random intercept/slope model with intercept variance and slope variance equal to $\tau_0^2 = 1$ and $\tau_1^2 = 1$, respectively. It can be seen that when there is no autocorrelation (i.e. $\rho = 0$), the optimal allocation of time points depends strongly on the covariance structure and priors and coinciding time points occur. Further, when the autocorrelation $\rho > 0$, the optimal allocations are never coinciding and are comparable for a prior variance equal to $\sigma^2 = 5$ and all covariance structures D. The effect of a large versus small autocorrelation is only presented for the fixed effects model ($\mathbf{D} = 0$), because Abebe et al. (2014c) already showed this for the random effects models. Finally, the prior means do not have much effect on the optimal allocation. This is in line with the findings of Abebe et al. (2014c) for a large prior variance.

Summary and Discussion

Optimal designs for longitudinal studies have been shown useful to improve the precision of the model parameter estimates of interest. Due to absence of a computer program for the optimal design of longitudinal studies with a binary response, planners of such studies in psychology, health sciences, and medicine have not yet benefitted from optimal design theory. We present a user-friendly computer program that computes Bayesian optimal designs for mixed effects logistic models with polynomial time effects. This computer program helps researchers to identify the optimal number and allocations of time points of measurements for a given subject-to-measurement cost ratio, and computes the loss in efficiency of equidistance compared to the optimal allocation. Moreover, it helps to assess the effect of autocorrelation on optimal allocations of design points. The program was illustrated on a smoking prevention study showing that, when the cost ratio k is less than 5, the optimal number of repeated measurements is 4 time points. Further, the optimal allocation is not equidistant, but equidistance is highly efficient.

The use of a Bayesian design does not force researchers to use Bayesian methods to analyze the data. Once the experimental data is collected by using the Bayesian D-optimal design, researchers can fit their model either with Bayesian or with frequentist methods.

The current version of the MATLAB program BODMixed_Logistic is freely available upon request from the corresponding author, which may be available eventually via the internet. The current version of the program considers designs based on the D-optimality criterion and assumes that all subjects are available over the total study period and that there is no dropout. Further, extensions of the model and software can be made by, e.g., adding a grouping variable or covariates like age or allowing for different types of covariance structures than already described in this paper. Future work may therefore aim at these extensions and at allowing for dropout. Another important issue for future work is Bayesian optimal design for model using non-polynomial (splines) time effects.

References

Abebe, H. T., Tan, F. E. S., van Breukelen, G. J. P., & Berger, M. P. F. (2014a). Bayesian *D*-optimal designs for the two parameter logistic mixed effects model. *Computational Statistics and Data Analysis, 71*, 1066-1076. doi: 10.1016/j.csda.2013.07.040

Abebe, H. T., Tan, F. E. S., van Breukelen, G. J. P., & Berger, M. P. F. (2014b). On the choice of a prior for Bayesian D-optimal designs for the logistic regression model with a single predictor. *Communications in Statistics - Simulation and Computation, 43*(7), 1811-1824. doi: 10.1080/03610918.2012.745556

Abebe, H. T., Tan, F. E. S., van Breukelen, G. J. P., & Berger, M. P. F. (2014c). Robustness of Bayesian D-optimal design for the logistic mixed model against misspecification of autocorrelation. *Computational Statistics, 29*(6), 1667-1690. doi: 10.1007/s00180-014-0512-3

Abebe, H. T., Tan, F. E. S., van Breukelen, G. J. P., & Berger, M. P. F. (2015). Bayesian design for dichotomous repeated measurements with autocorrelation. *Statistical Methods in Medical Research, 24*(5), 594-611. doi: 10.1177/0962280213508850

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. Oxford, UK: Clarendon Press.

Ausems, M., Mesters, I., van Breukelen, G., & De Vries, H. (2002). Short-term effects of a randomized computer-based out-of-school smoking prevention trial aimed at Dutch elementary schoolchildren. *Preventive Medicine, 34*(6), 581-589. doi: 10.1006/pmed.2002.1021

Berger, M. P. F. (1986). A comparison of efficiencies of longitudinal, mixed longitudinal, and cross-sectional designs. *Journal of Educational Statistics, 11*(3), 171-181. doi: 10.3102/10769986011003171

Berger, M. P. F., & Wong, W. K. (2009). *An introduction to optimal designs for social and biomedical research*. Chichester, UK: Wiley.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9-25. doi: 10.2307/2290687

Bunke, H, & Bunke, O. (1986). *Statistical inference in linear models* (Vol.1). Chichester, UK: Wiley.

Chaloner, K., & Larntz, K. (1989). Optimal Bayesian designs applied to logistic regression experiments. *Journal of Statistical Planning and Inference, 21*(2), 191-208. doi: 10.1016/0378-3758(89)90004-9

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science, 10*(3), 273-304. doi: 10.1214/ss/1177009939

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics, 24*(4), 586-602. doi: 10.1214/aoms/1177728915

Firth, D., & Hinde, J. P. (1997). On Bayesian D-optimum design criteria and the equivalence theorem in non-linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 59*(4), 793-797. doi: 10.1111/1467-9868.00096

Gilks, W. R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J. M Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4: Proceedings of the Fourth Valencia International Meeting, April 15-20, 1991* (pp.641-649). Oxford, UK: Clarendon Press.

Gilks, W. R., Best, N. G., & Tan, K. K. C. (1995). Adaptive rejection metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 44*(4), 455-472. doi: 10.2307/2986138

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 41*(2), 337-348. doi: 10.2307/2347565

Han, C., & Chaloner, K. (2004). Bayesian experimental designs for nonlinear mixed models with application to HIV dynamics. *Biometrics, 60*(1), 25-33. doi: 10.1111/j.0006-341X.2004.00148.x

Hartung, C., Willcutt, E., Lahey, B., Pelham, W., Loney, J., Stein, M., & Keenan, K. (2002). Sex differences in young children who meet criteria for attention deficit hyperactivity disorder. *Journal of Clinical Child & Adolescent Psychology, 31*(4), 453-464. doi: 10.1207/S15374424JCCP3104_5

Jang, W., & Lim, J. (2009). A numerical study of PQL estimation biases in generalized linear mixed models under heterogeneity of random effects. *Communication in Statistics – Simulation and Computation, 38*(4), 692-702. doi: 10.1080/03610910802627055

Lahey, B., Pelham, W., Stein, M., Loney, J., Irapani, C., Nugent, K.,…Baumann, B. (1998). Validity of DSMIV attention-deficit/hyperactivity disorder for younger children. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(7), 695-702. doi: 10.1097/00004583-199807000-00008

MathWorks. (2010). MATLAB (Version 7.10.0.499 (R2010a)) [computer software]. Natick, MA: The MathWorks, Inc.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2*(1), 3-19. doi: 10.1037/1082-989X.2.1.3

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society. Series D (The Statistician), 50*(1), 17-30. doi: 10.1111/1467-9884.00257

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2003). A comparison of estimation methods for multilevel logistic models. *Computational Statistics, 18*(1), 19-37. doi: 10.1007/s001800300130

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York, NY: Springer.

Niaparast, M. (2009). On optimal design for a Poisson regression model with random intercept. *Statistics & Probability Letters, 79*(6), 741-747. doi: 10.1016/j.spl.2008.10.035

Niaparast, M., & Schwabe, R. (2013). Optimal design for quasi-likelihood estimation in Poisson regression with random coefficients. *Journal of Statistical Planning and Inference, 143*(2), 296-306. doi: 10.1016/j.jspi.2012.07.009

Ouwens, M. J. N. M., Tan, F. E. S., & Berger, M. P. F. (2006). A maximin criterion for the logistic random intercept model with covariates. *Journal of Statistical Planning and Inference, 136*, 962-981. doi: 10.1016/j.jspi.2004.07.014

Raudenbush, S. W., & Feng, L. X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401. doi: 10.1037/1082-989X.6.4.387

Sebastiani, P., & Settimi, R. (1998). First-order optimal design for non-linear models. *Journal of Statistical Planning and Inference, 74*(1), 177-192.

Silvey, S. D. (1980). *Optimal design: An introduction to the theory for parameter estimation*. New York, NY: Chapman and Hall.

Tan, F. E. S. (2011). Conditions for D_A–maximin marginal designs for generalized linear mixed models to be uniform. *Communications in Statistics-Theory and Methods, 40*(2), 255-266. doi: 10.1080/03610920903411218

Tan, F. E. S., & Berger, M. P. F. (1999). Optimal allocation of time points for the random-effects model. *Communication in Statistics – Simulation and Computation, 26*(2), 517-540. doi: 10.1080/03610919908813563

Tekle, F. B., Tan, F. E. S., & Berger, M. P. F. (2008). Maximin D-optimal designs for binary longitudinal responses. *Computational Statistics & Data Analysis, 52*(12), 5253-5262. doi: 10.1016/j.csda.2008.04.037

Tekle, F. B., Tan, F. E. S., & Berger, M. P. F. (2009). Interactive computer program for optimal designs of longitudinal cohort studies. *Computer Methods and Programs in Biomedicine, 94*(2), 168-176. doi: 10.1016/j.cmpb.2008.11.002

Zeger, S. L., Liang, K-Y, & Albert, P. S (1988). Model for longitudinal data: a generalized estimating equation approach. *Biometrics, 44*(4), 1049-1060. doi: 10.2307/2531734

Appendix A: Derivation for the Relative Efficiency Equation (14)

To compare designs we compute their efficiencies using the concept of equivalent sample size (see Atkinson et al., 2007, p. 152; Berger & Wong, 2009, p. 37). Let $\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_s}\right)$ and $\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_q}\right)$ be the variance-covariance matrices of $\hat{\boldsymbol{\beta}}$ for the design ξ_s with s time points and the design ξ_q with q time points, respectively, and let N_s and N_q be the number of subjects for the design ξ_s and ξ_q, respectively. For the D-criterion and a given model with p parameters, the relative efficiency of design ξ_s compared to design ξ_q is given by:

$$\text{RE}\left(\xi_s;\xi_q\right)=\frac{N_s}{N_q}\left[\frac{\det\left\{\left[\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_s}\right)\right]^{-1}\right\}}{\det\left\{\left[\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_q}\right)\right]^{-1}\right\}}\right]^{\frac{1}{p}} \tag{15}$$

Where the two determinants in (15) are both based on one subject only, and the factor N_s/N_q takes into account the sample size per design.

This relative efficiency (15) can be rewritten as follows:

$$\begin{aligned}\text{RE}\left(\xi_s;\xi_q\right)&=\frac{N_s}{N_q}\exp\left\{\log\left\{\left[\frac{\det\left\{\left[\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_s}\right)\right]^{-1}\right\}}{\det\left\{\left[\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_q}\right)\right]^{-1}\right\}}\right]^{\frac{1}{p}}\right\}\right\}\\&=\frac{N_s}{N_q}\exp\left\{\frac{\log\det\left\{\left[\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_s}\right)\right]^{-1}\right\}-\log\det\left\{\left[\text{var}\left(\hat{\boldsymbol{\beta}}_{\xi_q}\right)\right]^{-1}\right\}}{p}\right\}\end{aligned} \tag{16}$$

Rewriting N_s and N_q in terms of cost ratio k and number of time points for the same total cost using the cost function equation (13), i.e.,

$$N_s = \frac{C}{C_2(k+s)} \quad \text{and} \quad N_q = \frac{C}{C_2(k+q)}$$

we obtain

$$\mathrm{RE}(\xi_s;\xi_q) = \frac{k+q}{k+s}\exp\left\{\frac{\log\det\left\{\left[\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\xi_s}\right)\right]^{-1}\right\} - \log\det\left\{\left[\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\xi_q}\right)\right]^{-1}\right\}}{p}\right\} \quad (17)$$

This relative efficiency (17) is for locally optimal design, i.e., for given parameter values. By generalizing this to Bayesian design, the RE of design ξ_s compared to design ξ_q with prior distribution π for $\boldsymbol{\beta}$ becomes as follows:

$$\mathrm{RE}(\xi_s;\xi_q(\pi)\,|\,\pi) = \frac{k+q}{k+s}\exp\left\{\frac{\mathrm{E}_{\boldsymbol{\beta}}\log\det\left\{\left[\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\xi_s}\right)\right]^{-1}\right\} - \mathrm{E}_{\boldsymbol{\beta}}\log\det\left\{\left[\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\xi_q}\right)\right]^{-1}\right\}}{p}\right\} \quad (18)$$

Thus, using the Bayesian D-optimality criterion (12), the RE will be:

$$\mathrm{RE}(\xi_s;\xi_q(\pi)\,|\,\pi) = \frac{k+q}{k+s}\left[\exp\left\{\frac{\phi_{\mathrm{D}}(\xi_s\,|\,\pi) - \phi_{\mathrm{D}}(\xi_1\,|\,\pi)}{p}\right\}\right] \quad (19)$$

When the ratio $(k + q)/(k + s)$ is one, that is, if either $q = s$ or the cost ratio k is very large, this relative efficiency (19) becomes the same as the relative efficiency given by Chaloner and Larntz (1989).

Appendix B: BODMixed_Logistic Manual

Introduction

Bayesian Optimal Design for Mixed effects Logistic models with polynomial time effect (BODMixed_Logistic) is graphical user interface software that computes optimal designs for longitudinal studies with a binary response. The program runs in a MATLAB (32-bit version 7.10.0499 (R2010a)) environment. In any case, the program works on a HP Compaq 8200 Elite PC with Windows 7 Enterprise and configuration i5-2400 CPU, 3.1 GHz, 4 GB RAM memory and 64-bit operating system or comparable systems.

To start the program:

1. Start Matlab.
2. Choose the option Window → Workspace → Current Folder and choose the directory where the software is located.
3. Choose Window → Command window and type BODMixed_Logistic (case sensitive) press the ↵ Enter key.

After starting the BODMixed_Logistic program, the user will find the main menu of the BODMixed_Logistic program as shown in Figure 4. There are five panels that will each be explained in turn. In this paper, a tutorial section is included which discusses the various decisions that the user has to make when using the program to find the most efficient design.

First Panel: Input Values of the Model

- *Choose model type*: The user can choose the degree of the polynomial of the mixed logistic model, i.e., a linear (which is the default value), quadratic or cubic model for the trend over time.
- *Variance-covariance parameters (D)*: The user will find a sub-menu to enter the input values for the variances and covariances (matrix **D**) of the random parameters. Figure 5 shows the sub-menu for a quadratic model. A fixed effects logistic model is obtained by setting all values in **D** to zero. The matrix **D** must be specified for each run, i.e. the values of the previous run are not saved.

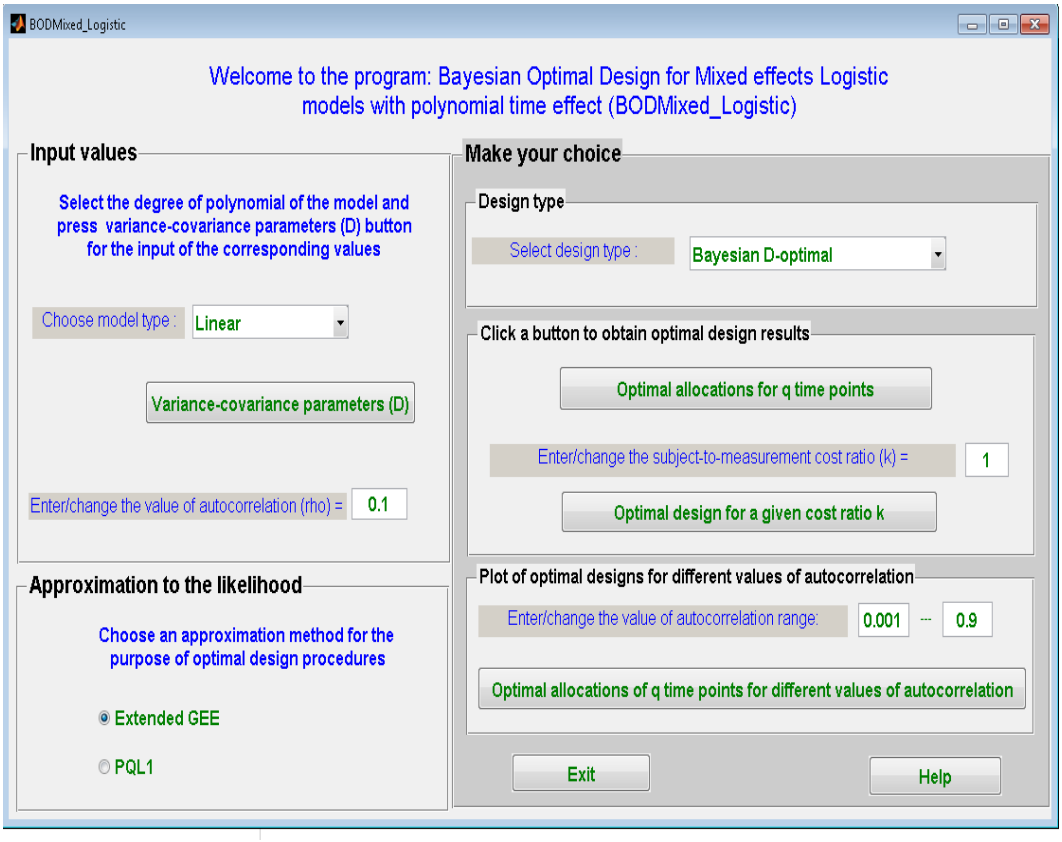


Figure 4. Layout of the main menu with the default input values for BODMixed_Logistic program

---

- *Enter/change the value of autocorrelation (rho)*: This is the size of the autocorrelation coefficient that the user expects between two repeated measurements at a time distance of one, i.e., $\rho = \text{Corr}(y_{ij}, y_{il})$ for which $|t_j - t_l| = 1$, keeping in mind that the total follow-up time is scaled to the interval [-1, +1] so that a time distance of 1 corresponds to half the follow-up time.

Second Panel: Computational Method

- *Approximation to the likelihood*: The user can choose an approximation method for the computation of optimal designs, i.e., either extended GEE or PQL1. The default method is the extended GEE.

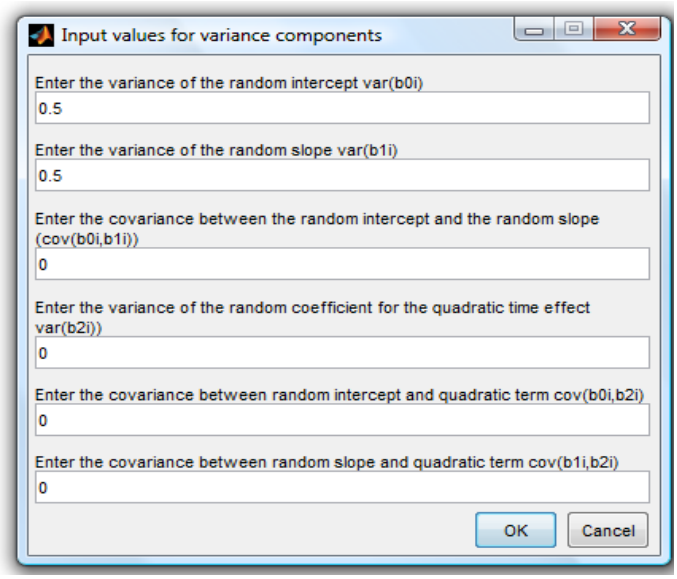


Figure 5. The sub-menu of BODMixed_Logistic for input values for variance components in the **D** matrix for the mixed logistic model with quadratic time effects

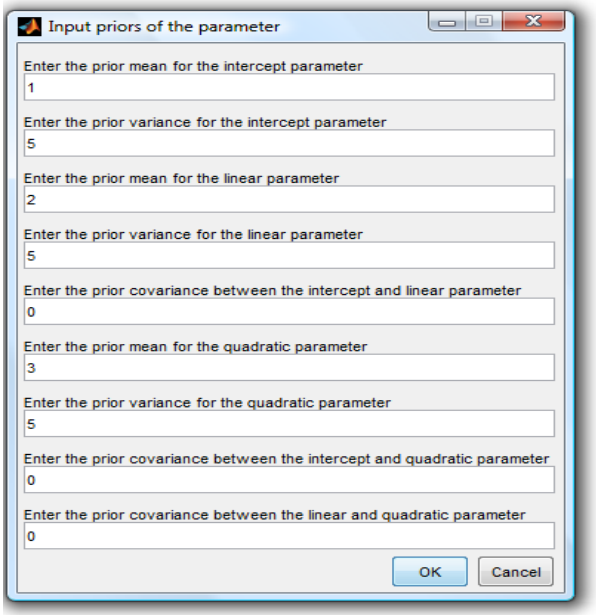


Figure 6. The sub-menu of BODMixed_Logistic for input values for the (normal) priors for the fixed effects parameters of the logistic model with quadratic time effect in the case of Bayesian design

Third Panel: Design Criterion

- *Select design type*: Either Bayesian D-optimal, locally D-optimal, or Maximin D-optimal design. When the user selects a design type, a sub-menu to fill in the input values for the relevant parameters will appear. Figure 6 is an example of a sub-menu for a Bayesian D-optimal design, where the prior means and prior variances can be specified. The input values must be filled in for each run, i.e. the values of the previous run are not saved.

Fourth Panel: Optimal Design Results

In this panel the user can choose between two methods of optimization:

- Fixing the number of time points at some value q to find the optimal allocation of those time points within the time interval [-1, +1],
- or fixing the subject-to-measurement cost ratio and letting the software then find the optimal number of time points as well as the optimal allocation.

- *Optimal allocations for q time points*: A dialog box appears to fill in a specific number of time points q (see Figure 7a). Then, the optimal allocations of time points within the time interval [-1, +1] will be found for the specified number of time points q, and the relative efficiency of equidistant time points compared to the optimal allocation will also be computed

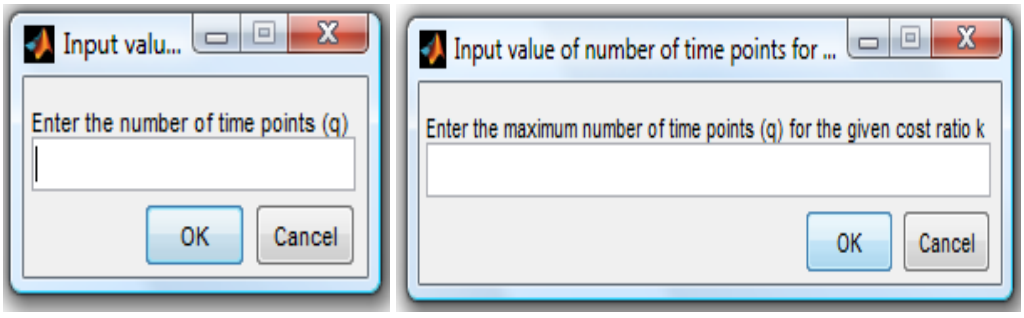


Figure 7. The sub-menu of BODMixed_Logistic to specify the (a) number of time points (q), left, and (b) maximum number of time points, right

- *Enter/change the subject-to-measurement cost ratio (k)*: This is the ratio of the cost of adding a new subject to the cost of an additional measurement per subject. This ratio is assumed to be greater than or equal to zero.
- *Optimal design for a given cost ratio k*: determines the optimal number (q) of repeated measurements as well as the optimal allocation of the q time points for a given subject-to-measurement cost ratio k. The user must specify the maximum allowable number of time points (see Figure 7b). Note that the minimum number of time points is two for a linear, three for a quadratic, and four for a cubic polynomial time effect model. These minima have been implemented in the program already.

Fifth Panel: Plot of Optimal Designs for Different Values of Autocorrelation

The optimal allocations within the time interval [-1, +1] for a given number of time points q can be computed for each autocorrelation value and plotted against the autocorrelation within the range chosen by the user.

- *Enter/change the value of autocorrelation range*: The user can enter a lower and upper bound for the autocorrelation parameter.
- Optimal allocations of q time points for different values of autocorrelation: The user gets the sub-menu of Figure 7a to choose the number of time points (q). Any value with $q \geq p$ can be filled in, where p is the number of fixed parameters of the model (p = 2, 3, or 4 for the linear, quadratic, or cubic model, respectively).

The user can change input values or obtain results by pressing the corresponding buttons on the main menu (Figure 4) as many times as he/she wishes. A ‘Help’ button is also available for guidance. The ‘Exit’ button in the main menu stops the program.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 722-742.
doi: 10.22237/jmasm/1493599080

ISSN 1538 − 9472

Genetic Algorithms for Cross-Calibration of Categorical Data

Suja M. Aboukhamseen
Kuwait University
Kuwait City, Kuwait

Rym A. M'Hallah
Kuwait University
Kuwait City, Kuwait

The probabilistic problem of cross-calibration of two categorical variables is addressed. A probabilistic forecast of the categorical variables is obtained based on a sample of observed data. This forecast is the output of a genetic algorithm based approach, which makes no assumption on the type of relationship between the two variables and applies a scoring rule to assess the fitness of the chromosomes. It converges to a good-quality point probability forecast of the joint distribution of the two variables. The proposed approach is applied both at stationary points in time and across time. Its performance is enhanced when additional sampled data is included, and can be designed with different scoring rules or made to account for missing data.

Keywords: categorical variables, cross-calibration, genetic algorithms, probability forecasting

Introduction

Estimating the joint probability distribution of two categorical variables, based on observed data, is a common yet elusive statistical problem. Depending on the nature of the categorical variables and the intricacies that characterize their relationship, such an endeavor can be highly technical and computationally intensive. In addition, the observed data used to estimate the relationship often contains numerous sources of error or bias. Such errors, generally due to operators, equipment, or the environment, further complicate the problem; impairing the validity of any inference.

Statistical calibration models the relationship between two variables that measure the same characteristic. It saves researchers, industrials and technicians valuable time, money and effort by providing a mechanism that gives a more accurate measurement to a corrupted reading (Osborne, 1991). Its application is

Suja M. Aboukhamseen is a lecturer and researcher. Email her at saboukhamseen@yahoo.com.

particularly vital in two cases. The first case emanates when the data consists of precise measurements acquired using an invasive, destructive, costly, or time-consuming technique. In such a situation, there usually exists an alternative measurement scheme that is more complaisant but not as reliable. Paired samples from the two measurements may be calibrated; thus, providing a mechanism to forecast the more reliable method from the less reliable one. The second case arises in problems requiring data comparability. It occurs when more than one technique gives valid and reliable measurements of a certain characteristic and there is a need for cross comparison, over time or across individuals. This cross comparison or mapping or translating of one measurement of a specific phenomenon to another is known as cross-calibration.

In both cases, the data may be quantitative or qualitative (categorical). The nature of categorical data brings its own set of challenges. The data may be self-reported or may consist of self-responses/assessments. The challenge herein lies in assessing the different ways individuals apply and interpret categorical response scales (Salomon et al., 2004; Murray et al., 2002; van Buuren & Hopman-Rock, 2001). However, the calibration of such variables requires that the mapping process be customized to fit the nature of their relationship. The traits that characterize the relationship must be explicitly stated in order to maintain its integrity during the translation process. Catering to the requirements of the statistical association often means imposing restrictions on the outcomes through complex mathematical models and structures.

Assume that X and Z are categorical random variables that measure the same qualitative random phenomenon with r and c possible classes, respectively. Let π be the matrix of joint probabilities of X and Z where $\pi_{ij} = P(X = i, Z = j)$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. Further assume that π is unknown, but that there exists an observed sample of N pairs of qualitative readings on (X, Z) of the single characteristic of interest. The N pairs are cross-classified into an $r \times c$ contingency table $\mathbf{n}$ which represents the observed relationship between the categories of the two variables X and Z. In the contingency table, the cell frequency n_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, denotes the number of readings classified simultaneously into category i by the qualitative reading on X and into category j by the qualitative reading on Z, with $\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = N$. Let the observed relative frequency distribution corresponding to the contingency table $\mathbf{n}$ be denoted by $\mathbf{p}$ where $p_{ij} = \frac{n_{ij}}{N}$, $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

The objective is to use the observed relative frequency distribution **p** to find an estimate of the functional translation π between X and Z; explicitly to estimate the conditional distributions $P(Z \mid X)$ and $P(X \mid Z)$. Since both distributions $P(Z \mid X)$ and $P(X \mid Z)$ are derived from the joint $P(X, Z)$, it is sufficient to find the joint probability function π. The notions behind the science of probability forecasting are used to derive an estimate of π.

DeGroot and Fienberg (1983), Dawid (1982), Schervish et al. (2014) and others established guidelines as to what constitutes a good forecasting generating system. However, how to construct that system remains an open question. In some fields, the forecasting mechanism relies heavily on expert opinion. In others, more objective procedures are employed. Herein, our focus is on the development of a forecasting generating system. A genetic algorithm (GA) -based method is applied, that searches for a (near-)optimal translation between the variables of interest. The translation corresponds to a joint distribution in the form of a probability forecasting system, from which predictive estimates of one of the two variables may be generated for a specific set of values of the other variable.

A primary advantage of this approach is that it obtains this translation without explicitly accounting for constraints that characterize the nature of the relationship between the variables. It uses the observed sample data to guide the search process. Specifically, the GA fitness construct, which is based on methods developed in probability forecasting theory (DeGroot and Fienberg, 1983; Lichtenstein et al., 1982; Gneiting and Katzfuss, 2014), ensures that the generated forecasts are valid and that they are the best among all forecasts in their class.

The purpose of this study, therefore, is to provide an overview of applications of cross calibration and genetic algorithms, and to propose a genetic algorithm. To further improve the reliability of the generated estimates, a quasi-Markov element is added to the analysis. It extends the method to cross-calibrate categorical variables measured longitudinally over time, where the calibration forecasts are generated both forward and backward on a time scale. Incorporating time broadens the applicability of the methodology. It models the relationship in a manner that is closer to the true state of nature, thus enhancing the accuracy of the estimates. This is supplemented with an illustrative example using stroke rehabilitation data.

Background

Applications of Cross-Calibration

The importance of cross-calibration emanates not only from the savings it induces, but mainly from its wide areas of applicability. Applications for this kind of analysis are manifold, making statistical calibration a valuable analytical tool. Possible fields of applications are demography, psychology, engineering, and item response theory. The following presents many fields requiring cross-calibration.

In surveys, cross-calibration facilitates the comparison of results from different questionnaires and the evaluation of response consistency. In corrosion analysis, a fundamental part of engineering, pipes and wires of oil fields are subject to corrosion because of harsh weather conditions. Following up the progress of corrosion is essential not only to production and transport of oil products but most importantly to the safety of the equipment and the personnel. Accurate tests for the state of corrosion are often invasive, destructive, and costly. The use of statistical calibration provides an efficient cost-effective alternative.

In the computation of official statistics, indicators are essential in monitoring and assessing the performance of a nation's public policy agenda, development, and how far a nation has come along in attaining its goals. Because the concepts stated above are intuitively understood, standards for their computation and compilation tend to vary widely depending on the country and the era. This makes the comparison of indicators either among countries or over time within countries exceedingly difficult. In light of today's United Nations' millennium goals, many nations are eager to show how far they have come towards attainment. This is only possible through valid data comparison, which is achievable via cross-calibration (Murray et al., 2002).

Similarly, in medicine, the assessment of a given treatment may be conducted differently depending on the researcher's preference or the time in which the study was carried out. The development of a quantitative translation between them enables the comparison of clinical trials in particular those requiring a longitudinal design over time (van Buuren et al., 2001).

In psychometrics, evaluating people's abilities, attitudes, and cognition through the process of testing and scoring is essential. Item response theory (IRT) is used in psychometrics to develop and refine tests that measure latent traits of individuals. The development of reliable techniques to measure traits such as intelligence and scholastic aptitude are of primary aim/essence of common exams, and tests of certification, such as the GRE and GMAT exams. Calibration is used

in IRT to provide a frame of reference to interpret test results, to equate tests, and to unify measurement scales both within the test items of a single test and between tests. The current practice in many of these applications is limited in scope. In some, such as IRT, the analysis requires impractical and unrealistic assumptions of independence between the items (categories) under investigation. Other applications require complicated models that tailor each aspect of the relationship separately and impose assumptions that are at many times invalid. As a result, the translations produced by the calibration model may be deficient and inaccurate. The proposed method overcomes these pitfalls by applying a methodology that makes no assumption on the type of relationship between the categorical variables under consideration.

GA Applications in Statistics

GAs mimic the role nature plays in refining and improving creation. GAs apply selective procreation and survival of the fittest to produce (near-)optimal solutions. They start from an arbitrary initial population consisting of a set of K chromosomes, where each chromosome k, $k = 1, \ldots, K$, acts as a representative solution to the problem. The population undergoes an iterative process of selection, crossover, mutation, and survival of the fittest to form future generations; thus, instigating an artificial evolutionary process. The algorithm iterates until it satisfies a stopping criterion, which can be a prefixed number of iterations without improvement (i.e., convergence of the fitness function), a time limit, or a preset number of generations, n_g.

Many fields of science, such as bio-informatics, computer science, genetics, operations research, economics, engineering, quality control and mathematics, have benefited from GA's straightforward yet efficient solution strategies. GAs identified (near-)optima to numerous practical problems with varying degrees of complexity. Sayed et al. (2009) show that GAs and their hybrids can improve the predictive performance of regression models. Chen et al. (2015) apply an adaptive GA to forecast the holiday daily tourist volume based on seasonal tendency. Huang et al. (2014) used GAs to assess the quality of a certain type of salted meat based on three quality indices whose values are inferred from a colorimetric sensor array. Stojanovic et al. (2013) apply a self-adjusting GA to model the behavior of dams. Liu et al. (2013) develop a real-time GA that forecasts water quality in river crab aquaculture. Nieto et al. (2013) forecast the presence of cyanotoxins in the Trasona water reservoir of Northern Spain via GAs.

Örkcü (2013) construct a hybrid GA to choose the minimal subset of explanatory variables of a multiple linear regression model. Wibowo and Desa (2012) employ GA in conjunction with kernel principal component analysis to predict the non-linear relationship between surface roughness resulting from milling processes and the milling machine parameters in the presence of multiple collinearity. Huang (2012) designs a support vector regression GA for stock selection. Ahn et al. (2012) use GAs to forecast the appraisal value of a real estate. Aydilek and Arslan (2013) identify missing values in data sets via GAs.

In the field of scientific calibration, GAs are applied to estimate model parameters and generate predictions (Vitkovský et al., 2000). However, the application of GAs to statistical calibration in general and to categorical cross-calibration in particular remains limited.

Procedure

Although the observed relative frequency distribution **p** is a valid statistical point estimate of π, the true joint probability function of X and Z, it may, in many instances, be biased or corrupted because it is subject to numerous sources of errors. To obtain an alternative point estimate of π based on the same observed sample frequency distribution **p**, a GA evolutionary procedure is applied for categorical data. Unlike most GAs, the proposed GA design does not require encoding the data and maintains the data's structural integrity throughout the execution of the algorithm.

Chromosome's Definition and Fitness

When considering unknown outcomes from categorical variables, a common tacit employed is the probability of occurrence in each category. When generated for a future event, this probability is a point probability forecast. If the probability of occurrence is evaluated for each forecast category, then the sum of the probabilities should equal one; constituting a probability forecasting system. Given the available information, a probability statement about the unknown outcome of a categorical variable can be calculated and its competency evaluated.

Of the numerous criteria that are available to assess probability forecasts, (i.e. validity, refinement, etc.), calibration and scoring rules defined on the probabilities and their subsequently observed outcomes are among the more prevalent methods (Dawid, 1982; Gneiting & Katzfuss, 2014). A scoring rule is

the squared error function in which scores for all the forecast probabilities are aggregated and averaged to evaluate the system's predictive performance.

Even though originally developed for subjective probability forecasting in the field of meteorology, subjective probability forecasting has a broad applicability and a wide range of applications. For instance, it can be applied to cross-calibration and incorporated into the proposed GA as follows. For our purposes, we regard the GA chromosome k in generation g as an expression/propagation of some objective forecasting system $\hat{\pi}_k^g$. In this regard, the chromosome forecasting performance may be assessed and compared with other chromosomes.

GA, which is sequential in nature, obtains K possible estimates of π at each iteration (or generation), $g = 1, \ldots, n_g$. The $r \times c$ relative frequency matrix for the two categorical variables X and Z, $\hat{\pi}_k^g$, for each chromosome k, $k = 1, \ldots, K$, of iteration g is a possible estimate of π. The relative frequency $\hat{\pi}_{ijk}^g = \frac{\hat{\eta}_{ijk}^g}{N}$, represents the k^{th} probability forecast of $P(X = i, Z = j)$ at iteration g, where $\hat{\eta}_{ijk}^g$ are realizations from the k^{th} proposed joint probability $\hat{\pi}_{ijk}^g = P(X = i, Z = j)$ at iteration g of the number of times $X = i$ and $Z = j$. The sum of all frequencies, $\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{\eta}_{ijk}^g$, which equals N, is independent of g and k. Thus, the sum of all relative frequencies, $\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{\pi}_{ijk}^g$, always equals 1.

The fitness of a chromosome depends on the fitness of its genes. It reflects how well-calibrated the forecast frequency $\hat{\pi}_{ijk}^g$ is in comparison to p_{ij}, the observed proportion of times that $X = i$ and $Z = j$ in the observed data. A probability forecast is considered well calibrated if $\hat{\pi}_{ijk}^g = p_{ij}$. The larger the discrepancy between the observed relative frequency and the forecast probability, the less-calibrated the gene. Hence, the chromosome fitness $F_k^g, k = 1, \ldots, K$, is gauged by the scoring rule

$$F_k^g = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\hat{\pi}_{ijk}^g - p_{ij}\right)^2,$$

which is the sum of the squared differences of the observed and forecast frequency. The chromosomes within the population are hitherto evaluated and ranked according to this criterion. The fitness function $F_k^g, k = 1, \ldots, K$ is a proper

scoring rule (Brier, 1950). Therefore, it ensures the sharpness and calibration of the probability forecasts of the selected chromosome.

GA's Design

The proposed GA's design follows. The initial population consists of K randomly generated chromosomes. Only the fittest $\frac{K}{2}$ chromosomes of the population are granted procreation or crossover privileges. The other least fit $\frac{K}{2}$ chromosomes are deemed too weak and, therefore, unworthy of mating.

Crossover combines the genes of two existing chromosomes to generate two offspring. First, two chromosomes are selected to become parents, $Parent_1$ and $Parent_2$. Second, two integers s_1 and s_2 are randomly generated from the discrete intervals [1, r] and [1, c], respectively. Third, the sub-matrix consisting of the first s_1 rows and the first s_2 columns is cut out of $Parent_1$ and positioned on the same location on $Parent_2$, thus producing $Child_1$. This new offspring consists of the intersection of the first s_1 rows and s_2 columns of $Parent_1$ and of all other entries of $Parent_2$. Simultaneously, a sub-matrix of the same size and location is removed from $Parent_1$ and inserted into $Parent_2$ in the same way, giving rise to a second offspring, $Child_2$. This latter has the reverse composition of $Child_1$ with the sub-matrix of its first s_1 rows and s_2 columns emanating from $Parent_2$ and the remaining entries from $Parent_1$. Figure 1 illustrates the crossover of $Parent_1$ and $Parent_2$ to produce two children $Child_1$ and $Child_2$. The chromosomes are 5×3 matrices; i.e., categorical variables X and Z have 5 and 3 classes, respectively. The crossover chooses the two integer numbers $s_1 = 3$ and $s_2 = 2$ from the discrete uniforms [1,5] and [1,3], respectively. The light grey shaded areas of the parent chromosomes combine to form $Child_1$ and the dark grey shaded areas constitute $Child_2$.

To preserve the uniformity and hence the coherence of the new offspring, the alleles within $Child_1$ and $Child_2$ must be re-scaled. This requires that the relative frequencies in the child add up to 1. This is done by dividing each relative frequency by the existing total. The offspring are then merged with the existing population of generation g which consists of the $\frac{K}{2}$ parents that were involved in crossover and the $\frac{K}{2}$ childless chromosomes. The merged population has $\frac{K^2}{2}$

chromosomes: the K chromosomes of generation g and the $2\frac{K}{2}\left(\frac{K}{2}-1\right)$ offspring chromosomes. The merged population is then assessed and ranked.

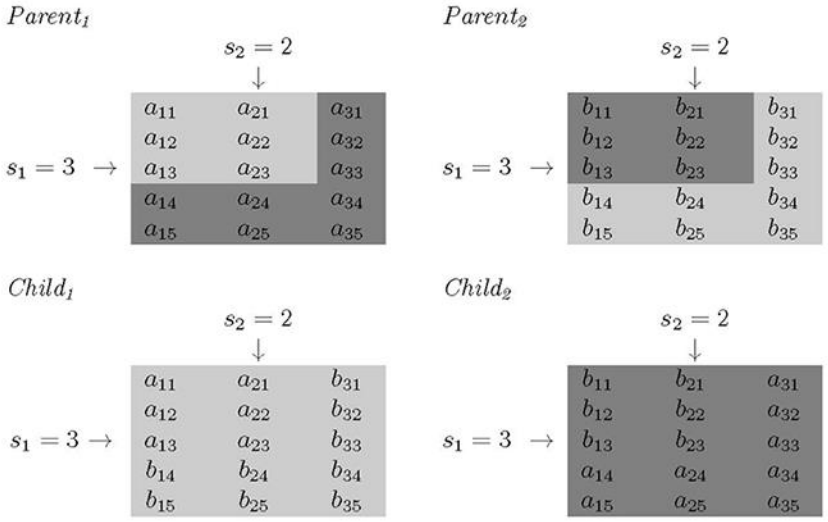


Figure 1. Crossover of two 5 × 3 parent chromosomes with $s_1 = 3$ and $s_2 = 2$ crossover points.

Further evolution of the population is enabled through mutation. For each chromosome $k, k = 1, \ldots, K$ in the population of generation g, a random probability measure $\alpha_k \in [0,1]$ is generated. If α_k is greater than α, the probability of mutation, the chromosome k is subject to a random swap of two of its alleles as follows. Two random integers s_1 and $s_{1'}$ (resp., s_2 and $s_{2'}$) are randomly chosen from the discrete uniform [1, r] (resp. [1, c]). The entries corresponding to $\hat{\pi}^{g}_{s_1 s_2 k}$ and $\hat{\pi}^{g}_{s_{1'} s_{2'} k}$ of k are then swapped. Mutation does not require the re-scaling of the alleles since the total relative frequency is fixed. The mutant replaces the least fit chromosome of the population if the former improves the latter. Once it completes the mutation step, GA ranks the population again.

To maintain the vitality of the population, GA culls the weakest chromosomes. Applying the survival of the fittest principle, GA selects the elite group consisting of the fittest K chromosomes of the mutated population. This group serves as the population of the next generation or iteration $g + 1$.

GA iterates through the above steps (i.e., crossover, mutation, and selection) until it satisfies a stopping criterion. Preliminary testing of the algorithm suggests that the stopping criterion should be a preset number of iterations $n_g = 1{,}000$. It ensures reasonably well-calibrated forecasts with a negligible fitness value of the best chromosome.

The above GA determines the joint probability distribution of two categorical variables X and Z based on an observed sample of paired observation. This distribution is used to determine the conditional probabilities of X given Z and of Z given X. However, the joint and conditional distributions are valid for a stationary point in time. In the following, the GA is extended to account for a time component (if applicable). Thus, GA will provide point probability forecasts for future or past points in time; allowing for the comparison of results of scientific studies undertaken at different points on the time horizon.

GA Across Time

For applications that involve time, GA is altered so that it evolves over time in a manner similar to a Markov chain. Let $t = t_1, t_2, t_3, \ldots$, represent sequential points in time. At any arbitrary initial point in time t_l, the GA is executed as described above until a well-calibrated population, $\mathbf{P}_{t_l}$, comes to term. To move either forward or backward to instant t_l, the GA is executed once more using P_{t_l} as the initial population. The transition in time is made possible by altering the fitness function to

$$F_k^{t_{l'}} = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\hat{\pi}_{ij}^{t_{l'}} - p_{ij}^{t_l}\right)^2 .$$

When applied forward (resp. backward) in time, this procedure sets $t_{l'}$ to t_{l+1} (resp. t_{l-1}). Time points do not need to be equally spaced on the time horizon. Explained in Figure 2 is the application of GA for transitions, where the present time is indicated via a dashed arrow and the future/past via a solid arrow. At the present time t_l, the initial population is generated randomly and GA is applied. The outcome of GA at the present time is then used as the initial population for the time $t_{l'}$, regardless of whether $t_{l'} = t_{l+1}$ or t_{l-1}.

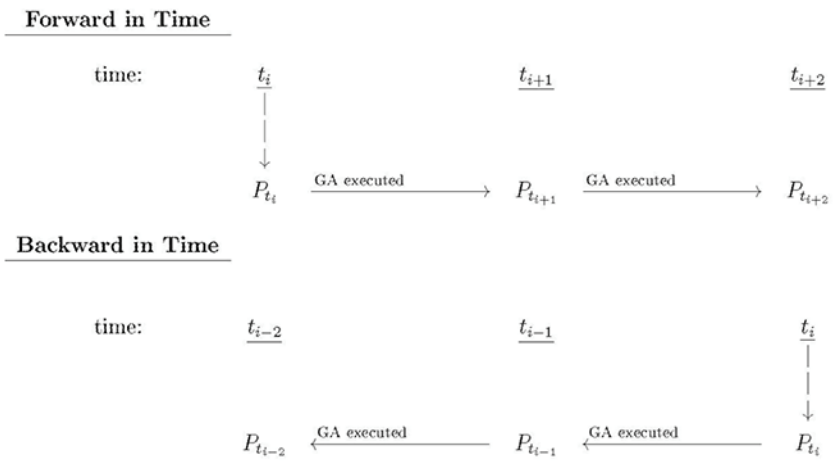


Figure 2. Forward and backward transition of GA in time.

A Cross-Calibration Application

In the assessment of stroke victims, standardized disability measures are commonly used. The scales are crucial in understanding the effectiveness of stroke treatments; yet, seldom is a patient assessed on more than one scale. A translation between two scales allows for the comparison among clinical trials and aids the development of alternative treatments.

Consider two commonly used standardized stroke disability measures, and apply GA cross calibration to form a feasible translation between them. The first is the Barthel Activity of Daily Living (ADL) Index (BI) attributed to Mahoney and Barthel (1965). It is a general measure of ADL, applied to a spectrum of medical conditions. The second is the Modified Rankin Outcome Scale (RS) (Rankin, 1957). It is a measure of the severity of disability in stroke victims. Currently, it is the most widely used measure of disability assessment for stroke victims (Saver et al., 2010). Much work has been done to compare the effectiveness of the measures and to determine whether the same clinical conclusion can be drawn from them (Sulter et al., 1999; Saver et al., 2010; Uyttenboogaart et al., 2007).

The BI defines 10 criteria of basic ADL and assesses the patients' capability to perform each of them. A minimum score of 0 is given if the patient is incapable of carrying out the task, and a maximum score is attributed if the patient can perform the ADL task independently. Partial scores, presented in increments of 5,

are allocated to patients who can perform the tasks, but with varying degrees of assistance. The scores of the 10 tasks are compiled to create an aggregate score with a maximum of 100. That is, a BI score of 100 indicates that the patient is physically independent.

The RS score assigns patients a discrete score from 0 to 5 depending on their degree of reliance on assistance and care. In contrast to the BI measure, a maximum RS score of 5 indicates the patient has severe disability and is highly dependent on nursing assistance. Whereas, a patient who exhibits no symptom of stroke debilitation and is independent is given a score of 0. Table 1 describes the 5 RS rankings and the 10 ADL criteria assessed by BI and their maximal achievable scores.

Table 1. The different measurement schemes: their measurement criteria and scores.

a. The BI criteria for ADL

Item	*Maximum score*
Feeding	10
Transferring	15
Grooming	5
Toileting	10
Bathing	5
Walking	15
Stairs	10
Dressing	10
Bowel continence	10
Bladder continence	10

b. The Modified Rankin Scale

Item	*Score*
No symptoms	0
No significant disability	1
Slight disability	2
Moderate disability	3
Moderately severe disability	4
Severe disability	5

The data used in this example was taken from the Kansas City Stroke Study (KCSS), a prospective cohort study of 459 individuals designed to characterize the patterns of recovery of patients with mild, moderate, and severe stroke. As described by Duncan et al. (2000), the 459 individuals with stroke were assessed using both the BI and RS instrumentations 14 days after the incidence of stroke. A follow-up was performed at 1, 3, and 6 months after stroke. Table 2 summarizes the observed data.

All data was collected from hospitals in the Greater Kansas City area. The rating of the stroke patients in the study was performed on both the RS and BI scales by either a physical therapist or a study nurse. Despite the fact that the same enumerator rated each patient, the data is still subject to numerous sources of measurement error. One possible source is the two groups of raters: the study

nurses and the physical therapists. There can be differences both between and within these two groups on how they perceive and interpret the disability criteria measures. Likewise, a stroke patient's subjective interpretation of daily functions can vary widely from patient to patient depending on a wide spectrum of factors such as the patient's level of activity pre and post the advent of stroke. Another source of measurement error is how the enumerator perceives the patients' activity and the many interaction effects therein. All of these factors (among others) culminate adding noise to the observed sample distorting the true distribution of the data.

Table 2. Cross tabulation of the ADL scores of the KCSS at 1, 3, and 6 months after the onset of a stroke. The columns represent the RS score. The rows are the BI.

Month 1							**Month 2**							**Month 3**						
BI\RS	***0***	***1***	***2***	***3***	***4***	***5***	***BI\RS***	***0***	***1***	***2***	***3***	***4***	***5***	***BI\RS***	***0***	***1***	***2***	***3***	***4***	***5***
0	0	0	0	0	1	10	*0*	0	0	0	0	1	7	0	0	0	0	0	0	5
5	0	0	0	0	0	9	*5*	0	0	0	0	0	2	5	0	0	0	0	1	2
10	0	0	0	0	3	1	*10*	0	0	0	0	0	2	10	0	0	0	0	1	1
15	0	0	0	0	2	1	*15*	0	0	0	0	5	0	15	0	0	0	0	3	0
20	0	0	0	0	5	3	*20*	0	0	0	0	3	0	20	0	0	0	0	3	2
25	0	0	0	0	7	1	*25*	0	0	0	0	7	0	25	0	0	0	0	4	1
30	0	0	0	0	7	0	*30*	0	0	0	0	6	0	30	0	0	0	0	4	0
35	0	0	0	0	8	0	*35*	0	0	0	0	7	0	35	0	0	0	1	3	0
40	0	0	0	2	14	0	*40*	0	0	0	0	3	0	40	0	0	0	0	4	0
45	0	0	0	0	4	0	*45*	0	0	0	0	3	0	45	0	0	0	0	2	0
50	0	0	0	1	8	0	*50*	0	0	0	1	6	0	50	0	0	0	2	3	0
55	0	0	0	1	9	0	*55*	0	0	0	0	5	0	55	0	0	0	3	5	0
60	0	0	1	5	9	0	*60*	0	0	0	3	6	1	60	0	0	0	3	4	0
65	0	0	1	5	3	0	*65*	0	0	0	4	3	0	65	0	0	1	0	5	0
70	0	0	1	12	3	0	*70*	0	0	1	11	8	0	70	0	0	0	6	1	0
75	0	0	1	19	6	0	*75*	0	0	0	5	0	0	75	0	0	0	12	2	0
80	0	0	1	18	3	0	*80*	0	0	6	12	0	0	80	0	0	0	9	0	0
85	0	0	4	26	0	0	*85*	0	2	4	21	0	0	85	0	0	3	18	0	0
90	1	0	7	24	1	0	*90*	1	2	9	23	0	0	90	0	3	11	13	0	0
95	1	4	31	13	0	0	*95*	1	4	24	20	0	0	95	2	6	35	16	0	0
100	2	17	62	11	0	0	*100*	7	44	72	9	0	0	100	11	57	62	11	0	0

Parmigiani et al. (2003) proposed a functional translation for the two measures using a statistical estimation approach. Although it produces adequate results, their approach requires that each characteristic of the relationship be modeled separately. GA avoids this. Its calibration accounts for all the relationship's characteristics intrinsically.

The objective is to determine the conditional probability distributions $P(BI|RS)$ and $P(RS|BI)$ at stationary time points and across time. Since both conditional distributions are functions of the joint distribution $P(BI,RS)$, GA determines only the latter. The GA is labeled vertical if applied at a stationary point in time and horizontal when applied either backward or forward across time.

Given in Table 3 are the joint distributions of RS and BI assessments at month 1. Table 3a is the result of a vertical GA at month 1 whereas Table 3b is the result of a backward GA starting at month 6 and moving in time to month 3 then to month 1. Both representations show good results; the negative correlation between the two scales is present, as expected, with higher probabilities attributed to the joint distribution of ratings along the counter diagonal in the lower triangle of Table 3.

Table 3. GA representations of the joint distributions after month 1 of the onset of a stroke. **a)** The joint distribution is independent of the information in months 3 and 6; **b)** The resulting joint distribution at month 1 when the GA is allowed to work backward in time from month 6 to month 3 to month 1.

a. Month 1: Random GA

BI\RS	*0*	*1*	*2*	*3*	*4*	*5*
0	0.000	0.000	0.000	0.000	0.006	0.027
5	0.000	0.000	0.000	0.000	0.000	0.025
10	0.000	0.000	0.000	0.000	0.008	0.006
15	0.000	0.000	0.000	0.000	0.006	0.006
20	0.000	0.000	0.000	0.000	0.013	0.008
25	0.000	0.000	0.000	0.000	0.020	0.006
30	0.000	0.000	0.000	0.000	0.019	0.000
35	0.000	0.000	0.000	0.000	0.024	0.000
40	0.000	0.000	0.000	0.006	0.049	0.000
45	0.000	0.000	0.000	0.000	0.012	0.000
50	0.000	0.000	0.000	0.006	0.023	0.000
55	0.000	0.000	0.000	0.006	0.026	0.000
60	0.000	0.000	0.006	0.013	0.024	0.000
65	0.000	0.000	0.006	0.013	0.008	0.000
70	0.000	0.000	0.006	0.032	0.008	0.000
75	0.000	0.000	0.006	0.051	0.018	0.000
80	0.000	0.000	0.006	0.049	0.008	0.000
85	0.000	0.000	0.011	0.067	0.000	0.000
90	0.006	0.000	0.020	0.078	0.007	0.000
95	0.006	0.011	0.067	0.049	0.000	0.000
100	0.006	0.018	0.067	0.033	0.000	0.000

b. Month 1: Time Reversal

BI\RS	*0*	*1*	*2*	*3*	*4*	*5*
0	0.000	0.000	0.000	0.000	0.000	0.020
5	0.000	0.000	0.000	0.000	0.000	0.011
10	0.000	0.000	0.000	0.000	0.000	0.011
15	0.000	0.000	0.000	0.000	0.014	0.000
20	0.000	0.000	0.000	0.000	0.011	0.000
25	0.000	0.000	0.000	0.000	0.019	0.000
30	0.000	0.000	0.000	0.000	0.017	0.000
35	0.000	0.000	0.000	0.000	0.020	0.000
40	0.000	0.000	0.000	0.000	0.010	0.000
45	0.000	0.000	0.000	0.000	0.011	0.000
50	0.000	0.000	0.000	0.010	0.020	0.000
55	0.000	0.000	0.000	0.000	0.014	0.000
60	0.000	0.000	0.000	0.011	0.021	0.000
65	0.000	0.000	0.000	0.012	0.010	0.000
70	0.000	0.000	0.011	0.031	0.022	0.000
75	0.000	0.000	0.000	0.014	0.000	0.000
80	0.000	0.000	0.019	0.043	0.000	0.000
85	0.000	0.011	0.011	0.063	0.000	0.000
90	0.010	0.011	0.023	0.064	0.000	0.000
95	0.011	0.011	0.071	0.064	0.000	0.000
100	0.021	0.133	0.094	0.024	0.000	0.000

Similarly good results are reported for month 3, as depicted in Table 4, which gives its joint distribution. These results were achieved by applying the GA

forward (Table 4a), backward (Table 4b), and vertically (Table 4c); thus, allowing for the comparison of the three probability forecasts at month 3. All three GA approaches perform well, but the retrospective GA provides the best results. This conclusion is based on the smallest value of the fitness function and on how well the joint distribution exhibits the nature of the relationship between RS and BI.

Table 4. GA representations of the joint distributions at month 3 after stroke onset. **a)** The joint distribution resulting from the GA going back in time from month 6 to month 3; **b)** The GA results independent of the information in months 3 and 6; **c)** The results of the GA moving forward in time from month 1 to month 3.

a. Time Reversal

BI\RS	*0*	*1*	*2*	*3*	*4*	*5*
0	0.000	0.000	0.000	0.000	0.000	0.020
5	0.000	0.000	0.000	0.000	0.000	0.008
10	0.000	0.000	0.000	0.000	0.000	0.008
15	0.000	0.000	0.000	0.000	0.012	0.000
20	0.000	0.000	0.000	0.000	0.008	0.000
25	0.000	0.000	0.000	0.000	0.020	0.000
30	0.000	0.000	0.000	0.000	0.020	0.000
35	0.000	0.000	0.000	0.000	0.023	0.000
40	0.000	0.000	0.000	0.000	0.010	0.000
45	0.000	0.000	0.000	0.000	0.008	0.000
50	0.000	0.000	0.000	0.008	0.020	0.000
55	0.000	0.000	0.000	0.000	0.020	0.000
60	0.000	0.000	0.000	0.010	0.023	0.000
65	0.000	0.000	0.000	0.011	0.008	0.000
70	0.000	0.000	0.000	0.031	0.021	0.000
75	0.000	0.000	0.000	0.020	0.000	0.000
80	0.000	0.000	0.020	0.038	0.000	0.000
85	0.000	0.010	0.011	0.061	0.000	0.000
90	0.000	0.011	0.025	0.064	0.000	0.000
95	0.000	0.009	0.084	0.068	0.000	0.000
100	0.020	0.124	0.118	0.027	0.000	0.000

b. Random GA

BI\RS	*0*	*1*	*2*	*3*	*4*	*5*
0	0.000	0.000	0.000	0.000	0.008	0.019
5	0.000	0.000	0.000	0.000	0.000	0.008
10	0.000	0.000	0.000	0.000	0.000	0.007
15	0.000	0.000	0.000	0.000	0.014	0.000
20	0.000	0.000	0.000	0.000	0.008	0.000
25	0.000	0.000	0.000	0.000	0.019	0.000
30	0.000	0.000	0.000	0.000	0.017	0.000
35	0.000	0.000	0.000	0.000	0.022	0.000
40	0.000	0.000	0.000	0.000	0.008	0.000
45	0.000	0.000	0.000	0.000	0.009	0.000
50	0.000	0.000	0.000	0.000	0.017	0.000
55	0.000	0.000	0.000	0.000	0.014	0.000
60	0.000	0.000	0.000	0.008	0.017	0.000
65	0.000	0.000	0.000	0.010	0.008	0.000
70	0.000	0.000	0.007	0.038	0.028	0.000
75	0.000	0.000	0.000	0.014	0.000	0.000
80	0.000	0.000	0.017	0.038	0.000	0.000
85	0.000	0.007	0.011	0.059	0.000	0.000
90	0.008	0.008	0.040	0.064	0.000	0.000
95	0.000	0.011	0.093	0.055	0.000	0.000
100	0.011	0.122	0.128	0.027	0.000	0.000

c. Forward in Time

BI\RS	*0*	*1*	*2*	*3*	*4*	*5*
0	0.000	0.000	0.000	0.000	0.006	0.022
5	0.000	0.000	0.000	0.000	0.000	0.007
10	0.000	0.000	0.000	0.000	0.000	0.007
15	0.000	0.000	0.000	0.000	0.025	0.000
20	0.000	0.000	0.000	0.000	0.008	0.000
25	0.000	0.000	0.000	0.000	0.020	0.000
30	0.000	0.000	0.000	0.000	0.019	0.000
35	0.000	0.000	0.000	0.000	0.020	0.000
40	0.000	0.000	0.000	0.000	0.007	0.000
45	0.000	0.000	0.000	0.000	0.009	0.000
50	0.000	0.000	0.000	0.006	0.021	0.000
55	0.000	0.000	0.000	0.000	0.020	0.000
60	0.000	0.000	0.000	0.017	0.020	0.007
65	0.000	0.000	0.000	0.011	0.009	0.000
70	0.000	0.000	0.007	0.033	0.023	0.000
75	0.000	0.000	0.000	0.022	0.000	0.000
80	0.000	0.000	0.017	0.050	0.000	0.000
85	0.000	0.010	0.011	0.064	0.000	0.000
90	0.008	0.007	0.025	0.079	0.000	0.000
95	0.000	0.017	0.074	0.063	0.000	0.000
100	0.025	0.045	0.118	0.042	0.000	0.000

Table 5 provides the joint distribution of RS and BI for month 6. Obtained in Table 5a is this joint distribution using the past information in month 1; a vertical GA is then applied for month 3 first then for month 6. Applied in Table 5b is a horizontal GA at month 6, using none of the data observed during months 1 and 6. Again, although both techniques show good results, the forward GA

produces slightly better results as it uses additional sample information for its forecast.

Table 5. GA representations of the joint distributions at month 6 after stroke onset. a. The resulting joint distribution at month 6 when the GA is allowed to move forward in time from month 1 to month 3 to month 6. b. The joint distribution independent of the information in months 1 and 3.

a. Month 6: Time Dependent

B\RS	*0*	*1*	*2*	*3*	*4*	*5*
0	0.000	0.000	0.000	0.000	0.006	0.027
5	0.000	0.000	0.000	0.000	0.000	0.025
10	0.000	0.000	0.000	0.000	0.008	0.006
15	0.000	0.000	0.000	0.000	0.006	0.006
20	0.000	0.000	0.000	0.000	0.013	0.008
25	0.000	0.000	0.000	0.000	0.020	0.006
30	0.000	0.000	0.000	0.000	0.019	0.000
35	0.000	0.000	0.000	0.000	0.024	0.000
40	0.000	0.000	0.000	0.006	0.049	0.000
45	0.000	0.000	0.000	0.000	0.012	0.000
50	0.000	0.000	0.000	0.006	0.023	0.000
55	0.000	0.000	0.000	0.006	0.026	0.000
60	0.000	0.000	0.006	0.013	0.024	0.000
65	0.000	0.000	0.006	0.013	0.008	0.000
70	0.000	0.000	0.006	0.032	0.008	0.000
75	0.000	0.000	0.006	0.051	0.018	0.000
80	0.000	0.000	0.006	0.049	0.008	0.000
85	0.000	0.000	0.011	0.067	0.000	0.000
90	0.006	0.000	0.020	0.078	0.007	0.000
95	0.006	0.011	0.067	0.049	0.000	0.000
100	0.006	0.018	0.067	0.033	0.000	0.000

b. Month 6: Random GA

B\RS	*0*	*1*	*2*	*3*	*4*	*5*
0	0.000	0.000	0.000	0.000	0.000	0.020
5	0.000	0.000	0.000	0.000	0.000	0.011
10	0.000	0.000	0.000	0.000	0.000	0.011
15	0.000	0.000	0.000	0.000	0.014	0.000
20	0.000	0.000	0.000	0.000	0.011	0.000
25	0.000	0.000	0.000	0.000	0.019	0.000
30	0.000	0.000	0.000	0.000	0.017	0.000
35	0.000	0.000	0.000	0.000	0.020	0.000
40	0.000	0.000	0.000	0.000	0.010	0.000
45	0.000	0.000	0.000	0.000	0.011	0.000
50	0.000	0.000	0.000	0.010	0.020	0.000
55	0.000	0.000	0.000	0.000	0.014	0.000
60	0.000	0.000	0.000	0.011	0.021	0.000
65	0.000	0.000	0.000	0.012	0.010	0.000
70	0.000	0.000	0.011	0.031	0.022	0.000
75	0.000	0.000	0.000	0.014	0.000	0.000
80	0.000	0.000	0.019	0.043	0.000	0.000
85	0.000	0.011	0.011	0.063	0.000	0.000
90	0.010	0.011	0.023	0.064	0.000	0.000
95	0.011	0.011	0.071	0.064	0.000	0.000
100	0.021	0.133	0.094	0.024	0.000	0.000

In all executions of GA for this example, the fitness function converges quickly. Despite its small magnitude, the fitness function never converges to zero. This only reiterates the fact that the observed sample data used for the assessment of the chromosomes' fitness contains some noise, the sources of which were enumerated earlier. Figure 3 demonstrates how the fitness function decreases with the population evolution at a stationary point in time.

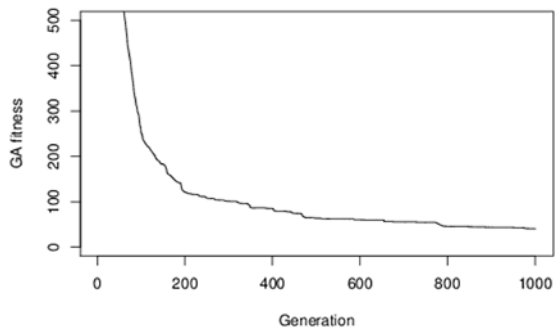


Figure 3. Convergence of the fitness function as the number of generations increases for a vertical GA applied at the stationary point 3 months after stroke onset.

In summary, GA performs well. The population converges quickly. In each generation, the chromosomes of the population display the negative correlations and properties that characterize the nature of the relationship between RS and BI. The time dependent GA performs better than the vertical GA because of the additional observed information being used. In particular, it is evident from the joint distribution of the first two time periods that a backward transition in time produces results that are more compliant with the expected nature of the relationship between the RS and BI measures.

Conclusion

Estimating the joint distribution of two categorical variables based on an observed sample data that contains some bias is an important topic and a cross-calibration problem. Because of its theoretical complexity and its widespread applications in several fields ranging from engineering to medicine to meteorology to population statistics. It is, herein, approximately solved using a non-traditional statistical method: genetic algorithm. Unlike other existing statistical methods, the adopted genetic algorithm does not make any assumption on the type or strength of the relationship between the categorical variables. It uses the observed sample to gauge the chromosomes of the successive populations. It converges rapidly to a good estimate of the true joint distribution. When applied over a time horizon, the genetic algorithm further enhances its estimates as it uses more observed data. When applied to the data collected for the Kansas City Stroke Study, it obtains logical point probability forecasts that concord with the true state of nature.

The proposed genetic algorithm based cross calibration approach can be tested with more sophisticated scoring rules or different fitness functions. Similarly, it can be applied to overcome missing data; in particular in clinical studies where subjects may move to different cities, die, or simply decide to stop participating in the study, and also in engineering set ups where the more reliable measurement methods are destructive or expensive.

References

Ahn, J., Byun, H., Oh, K., and Kim, T. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications, 39*(9), 8369–8379. doi: 10.1016/j.eswa.2012.01.183

Aydilek, A. and Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences, 233*, 25–35. doi: 10.1016/j.ins.2013.01.021

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Chen, R., Liang, C., Hong, W., and Gu, D. (2015). Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. *Applied Soft Computing, 26*, 435–443. doi: 10.1016/j.asoc.2014.10.022

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association, 77*(379), 605–610. doi: 10.1080/01621459.1982.10477856

DeGroot, M. and Fienberg, S. (1983). The comparison and evaluation of forecasters. *The Statistician, 32*(1), 12–22. doi: 10.2307/2987588

Duncan, P., Lai, S., and Keighley, J. (2000). Defining post-stroke recovery: implications for design and interpretation of drug trials. *Neuropharmacology, 39*(5), 835–841. doi: 10.1016/s0028-3908(00)00003-4

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application, 1*(1), 125–151. doi: 10.1146/annurev-statistics-062713-085831

Huang, C. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing, 12*(2), 807–818. doi: 10.1016/j.asoc.2011.10.009

Huang, X., Zou, X., Zhao, J., Shi, J., Zhang, X., Li, Z., and Shen, L. (2014). Sensing the quality parameters of Chinese traditional Yao-meat by using a colorimetric sensor combined with genetic algorithm partial least squares regression. *Meat Science, 98*(2), 203–210. doi: 10.1016/j.meatsci.2014.05.033

Lichtenstein, S., Fischhoff, B., and Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky, Eds. *Judgment Under Uncertainty: Heuristics and Biases*, pp. 306–334. Cambridge, England: Cambridge University Press. doi: 10.1017/cbo9780511809477.023

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., and Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Mathematical and Computer Modelling, 58*(3–4), 458–465. doi: 10.1016/j.mcm.2011.11.021

Mahoney, F. and Barthel, D. (1965). Functional Evaluation: The Barthel Index. *Maryland Medical Journal, 14*, 61–65.

Murray, C. J. L., Tandon, A., Salomon, J., Mathers, C., and Sadana, R. (2002). Cross-population comparability of evidence for health policy. In C. J. L. Murray & D. B. Evans, Eds. *Health Systems Performance Assessment: Debates, Methods and Empiricism*, pp. 705–713. Geneva, Switzerland: World Health Organization.

Nieto, P., Fernandez, J., de Cos Juez, F., Lasheras, F., and Muniz, C. (2013). Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (Northern Spain). *Environmental Research, 122*, 1–10. doi: 10.1016/j.envres.2013.01.001

Örkcü, H. H. (2013). Subset selection in multiple linear regression models: A hybrid of genetic and simulated annealing algorithms. *Applied Mathematics and Computation, 219*(23), 11018–11028. doi: 10.1016/j.amc.2013.05.016

Osborne, C. (1991). Statistical calibration: A review. *International Statistical Review, 59*(3), 309–336. doi: 10.2307/1403690

Parmigiani, G., Ashih, H., Samsa, G., Duncan, P. W., Lai, S., and Matchar, D. (2003). Cross-calibration of stroke disability measures. *Journal of the American Statistical Association, 98*(462), 273–281. doi: 10.1198/016214503000044

Rankin, J. (1957). Cerebral vascular accidents in patients over the age of 60: II. Prognosis. *Scottish Medical Journal, 2*(5), 200–215. doi: 10.1177/003693305700200504

Salomon, J., Tandon, A., and Murray, C. (2004). Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ, 328*(7434), 258–263. doi: 10.1136/bmj.37963.691632.44

Saver, J., Filip, B., Hamilton, S., Yanes, A., Craig, S., Cho, M., Conwit, R., and Starkman, S. (2010). Improving the reliability of stroke disability grading in clinical trials and clinical practice: The Rankin Focused Assessment (RFA). *Stroke, 41*(5), 992–995. doi: 10.1161/strokeaha.109.571364

Sayed, H., Gabbar, H., and Miyazaki, S. (2009). A hybrid statistical genetic-based demand forecasting expert system. *Expert Systems with Applications, 36*(9), 11662–11670. doi: 10.1016/j.eswa.2009.03.014

Schervish, M., Seidenfeld, T., and Kadane, J. (2014). Dominating countably many forecasts. *The Annals of Statistics, 42*(2), 728–756. doi: 10.1214/14-aos1203

Stojanovic, B., Milivojevic, M., Ivanovic, M., Milivojevic, N., and Divac, D. (2013). Adaptive system for dam behavior modeling based on linear regression and genetic algorithms. *Advances in Engineering Software, 65*, 182–190. doi: 10.1016/j.advengsoft.2013.06.019

Sulter, G., Steen, C., and De Keyser, J. (1999). Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke, 30*(8), 1538–1541. doi: 10.1161/01.str.30.8.1538

Uyttenboogaart, M., Luijckx, G., Vroomen, P., Stewart, R., and De Keyser, J. (2007). Measuring disability in stroke: relationship between the modified Rankin scale and the Barthel index. *Journal of Neurology, 254*(8), 1113–1117. doi: 10.1007/s00415-007-0646-0

van Buuren, S. and Hopman-Rock, M. (2001). Revision of the ICIDH severity of disabilities scale by data linking and item response theory. *Statistics in Medicine, 20*(7), 1061–1076. doi: 10.1002/sim.723

van Buuren, S., Eyres, S., Tennant, A., and Hopman-Rock, M. (2001). *Response conversion: A New Technology for Comparing Existing Health Information*. TNO Prevention and Health, Division Public Health.

Vitkovský, J., Simpson, A., and Lambert, M. (2000). Leak detection and calibration using transients and genetic algorithms. *Journal of Water Resources Planning and Management, 126*(4), 262–265. doi: 10.1061/(asce)0733-9496(2000)126:4(262)

Wibowo, A. and Desa, M. (2012). Kernel based regression and genetic algorithms for estimating cutting conditions of surface roughness in end milling

machining process. *Expert Systems with Applications, 39*(14), 11634–11641. doi: 10.1016/j.eswa.2012.04.004

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 743.

ISSN 1538 – 9472

Statistical Software Applications and Review

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 744-752.
doi: 10.22237/jmasm/1493599140

ISSN 1538 − 9472

Using Multiple Imputation to Address Missing Values of Hierarchical Data

Yujia Zhang
Centers for Disease Control and Prevention, Atlanta, GA

Sara Crawford
Centers for Disease Control and Prevention, Atlanta, GA

Sheree Boulet
Centers for Disease Control and Prevention, Atlanta, GA

Michael Monsour
Centers for Disease Control and Prevention, Atlanta, GA

Bruce Cohen
MA Department of Health
Boston, MA

Patricia McKane
MI Dept. of Comm. Health
Lansing, MI

Karen Freeman
FL Department of Health
Tallahassee, FL

Missing data may be a concern for data analysis. If it has a hierarchical or nested structure, the SUDAAN package can be used for multiple imputation. This is illustrated with birth certificate data that was linked to the Centers for Disease Control and Prevention's National Assisted Reproductive Technology Surveillance System database. The Cox-Iannacchione weighted sequential hot deck method was used to conduct multiple imputation for missing/unknown values of covariates in a logistic model.

Keywords: Hierarchical or nesting structure, multiple imputation, weighted sequential hot deck

Introduction

Population-based hierarchical or nested data and multiple covariates are often used in maternal and child health research. The covariates may contain unknown/missing values, which are excluded in traditional model fitting such that only complete cases are used. Although the percent of unknown/missing values for one variable is usually small, the percent of unknown/missing values across all covariates may be larger. Using only complete cases in analysis reduces the effective sample size and testing power, which is especially concerning when the

Yujia Zhang is a Mathematic Statistician in the Division of Reproductive Health, National Center for Chronic Disease Prevention and Health Promotion, CDC. Email at coi8@cdc.gov.

outcome is infrequent since it likely introduces small-sample bias in logistic model fitting (King & Zeng, 2001; Rotnitzky & Wypij, 1994).

One strategy to address the impact of missing values on parameter estimates is to use imputed data in analysis. A single imputation method fills each missing entry with an imputed value, such that standard complete-data methods can be used for analysis. This method ignores the variability contributed by the lack of information on the missing values, leading to variance underestimation. Another method, multiple imputation replaces each missing entry with two or more values and draws inferences by combining the results of several complete-data analyses to address within and between-imputation variability in variance estimation (Rubin, 1986, 1997; Schafer, 1999).

The traditional multiple imputation method used by most commercial statistical software packages such as SAS, IVEware, etc., adopts a parametric approach such as regression imputation modeling and imputes data under an assumption that the data follow a multivariate normal distribution. The multivariate normal distributional assumption may not always hold, especially for multilevel hierarchical data with very small clusters. The aim of the present study is to demonstrate a method of multiply imputing missing values for data with a hierarchical or nested data structure using a well-known statistical software package. This approach is demonstrated using SUDAAN's HOTDECK procedure (SUDAAN Release 11, RTI International, Research Triangle Park, North Carolina) and then fit logistic models using the multiply imputed data.

Data

A population-based dataset collected from multiple sources was used. It included live birth records (2000-2006) from Florida, Massachusetts, and Michigan linked to the National Assisted Reproductive Technology (ART) Surveillance System (NASS) at the Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention, 2014). The population of interest was infants conceived via ART. To eliminate the potential impact of subsequent treatments on maternal complications and pregnancy outcomes, only the first live born infant of the first live birth was included if a woman was identified as having more than one birth in the time period (Grigorescu, et al., 2014). Because the NASS data were reported by each fertility clinic in the United States, the data had a hierarchical structure and observations were nested in fertility clinics.

The main outcome of interest for our analysis was an Apgar score at five minutes, a binary variable corded as 0 (>=7) and 1 (<7). The Apgar score at five

minutes is the first test given to a newborn to quickly evaluate a newborn's physical condition with a score ranging from one to ten. Values of 7 and above are considered normal. The independent covariates in a logistic model were reason for ART (V_1), maternal age (V_2), race/ethnicity (V_3), education (V_4), adequacy of prenatal care (V_5), co-morbid conditions (V_6), delivery method (V_7), induction of labor (V_8), gestational age (V_9), newborn gender (V_{10}), and birth weight (V_{11}) (Grigorescu, et al., 2014).

Missing Value Imputation

SUDAAN was developed to analyze data from complex surveys; however SUDAAN is also able to analyze other hierarchical or nested data, or non-survey data. Data inspection showed that the amount of data missing for the outcome value was extremely small (<0.3%) so observations with missing outcome values were excluded, and imputed values only for observations with missing values for the covariates. SUDAAN's `HOTDECK` procedure was used to impute missing values of covariates, because 8.3% of the observations had a missing value for at least one covariate, resulting in a reduction of 67 cases. `HOTDECK` replaces missing values of one or more variables of a recipient using observed values from a "similar" respondent. Since our data were naturally clustered, i.e., the observations (infants) were clustered in fertility clinics, we restricted to obtaining the pool of respondents by clinic and replacing missing values of recipients in the same clinic. For each infant with missing values of the covariates (V_1, V_2, …, V_{11}), the HOTDECK procedure collected a set of similar infants from the same clinic (cluster) without missing covariates. From this set, randomly chosen infants were used to fill in the missing values of the covariates with replacement where each variable was filled separately. This process was repeated until all infants with missing values for covariates within the clinic were imputed. SUDAAN's `HOTDECK` procedure uses a weighted sequential hot deck method proposed by Cox (1980) and Iannacchione (1982) to perform imputation, the default method for `PROC HOTDECK`.

The SAS-callable SUDAAN was used with the following code for the `HOTDECK` procedure:

```
PROC HOTDECK DATA=DATA_INPUT SEED=3123845;
       IMPBY CLINIC;
       IMPID INFANT_ID;
       IMPVAR V1 V2 … V11/MULTIMP=5;
```

```
        WEIGHT _ONE_;
        IMPNAME V1="V1_IMP" V2="V2_IMP" … V11 = "V11_IMP";
        IDVAR APGAR;
        OUTPUT /IMPUTE=default FILENAME=OUTDATA REPLACE;
RUN;
```

In the PROC HOTDECK statement, DATA= specifies the input dataset (DATA_INPUT) which includes variables with missing values. The SEED= specifies an integer to generate a random number for the imputation. The cluster variable is specified on the IMPBY statement (CLINIC); data must be sorted by this cluster variable prior to running this procedure. Each observation clustered within the clinic is identified using the IMPID statement, in this case by the infant variable (INFANT_ID). The variables with missing values to be imputed (V_1, V_2, …, V_{11}) are listed in the IMPVAR statement. The option, MULTIMP=5, in the IMPVAR statement specifies that five imputed datasets are to be created. For the non-survey data, set the variable in the WEIGHT statement to be _ONE_, a default option in SUDAAN to indicate no weighting.

The IMPNAME statement assigns variable names for imputed variables (original variable name + IMP in our case). For each imputation, SUDAAN assigns a consecutive number after the imputed variable name (V1_IMP1 V2_IMP1 … V11_IMP1 in the first imputation, V1_IMP2 V2_IMP2 … V11_IMP2 in the second imputation, etc.). The IDVAR statement specifies that our outcome variable (APGAR), which was not imputed, should be included in the output dataset. The OUTPUT statement provides a dataset with all imputed variables, the cluster variable (specified by IMPBY), the imputation identification variable (specified by IMPID), and variables not imputed (specified by IDVAR). The option IMPUTE=default indicates that the output dataset will include all imputed variables ($11 \times 5 = 55$ imputed variables), the option FILENAME= specifies the name of the output dataset (OUTDATA), and the option REPLACE instructs SUDAAN to overwrite any existing dataset with the same name.

PROC MI in SAS (SAS v. 9.3, Cary, NC) was used to impute missing values in order to compare imputation results from PROC MI to those obtained from SUDAAN's PROC HOTDECK. The MI procedure is a parametric multiple imputation procedure that creates multiply imputed data sets using predicted values rather than observed values as HOTDECK to replace missing values. Due to some clinics having fewer than three observations (38.8% of total included clinics), PROC MI failed to provide any output for imputation. This demonstrates that the parametric imputation approach, such as sequential regression models, is limited in dealing

with very small clusters for multiple imputation. Because the `MI` procedure does not adequately perform imputation for the data, this method is not described in detail.

Statistical Analysis

Multiply imputed data was used. According to Rubin (1978), the multiple imputation estimator (denoted as $\hat{\theta}$) of parameter is the average of the estimators obtained from all K imputed datasets:

$$\bar{\theta}_K = \frac{1}{K}\sum_{i=1}^{K}\hat{\theta}_i, \tag{1}$$

The variance of $\bar{\theta}_K$ is the sum of the average within (imputed dataset)-imputation variance and the between (imputed datasets)-imputation variance. Because the population data was used, the finite population correction can be ignored, denoting the variance of the i^{th} imputed dataset as W_i, the average within-imputation variance is:

$$\bar{W}_K = \tfrac{1}{K}\sum_{i=1}^{K} W_i \tag{2}$$

and the between-imputation variance is:

$$B_K = \frac{1}{K-1}\sum_{i=1}^{K}\left(\hat{\theta}_i - \bar{\theta}\right)^2 \tag{3}$$

The overall variance of $\bar{\theta}_K$ is the sum of within-imputation variance and the between-imputation variance, with a bias correction for the finite number of multiply imputed data sets:

$$Var\left(\bar{\theta}_K\right) = \bar{W}_K + \tfrac{K+1}{K} B_K \tag{4}$$

The SAS-callable SUDAAN `RLOGISTIC` procedure was used to fit a random effects logistic regression model using imputed data. Collinearity was inspected between covariates using Zack's SAS Macro (n.d.) for the logistic model with the following `RLOGISTIC` procedure:

```
PROC RLOGIST DESIGN=WR DATA=IMPN1 MI_COUNT=5;
       NEST _ONE_ CLINIC;
       WEIGHT _ONE_;
       CLASS V1_IMP …;
       REFLEVEL V1_IMP=1 …;
       MODEL APGAR= V1_IMP V2_IMP … V11_IMP;
RUN;
```

In the PROC RLOGISTIC statement, set DESIGN = WR (sampling with replacement for population data, SUDAAN's default design). Using the output dataset from the imputation procedure (OUTDATA), we created 5 datasets (Sinharay, Stern, and Russell, 2001), one for each imputation, and each dataset included 14 variables, INFANT_ID, CLINIC, APGAR, V1_IMP, V2_IMP, …, V11_IMP for model fitting. Assign the names IMPN1, IMPN2, IMPN3, IMPN4 and IMPN5 to these datasets. The options DATA=IMPN1 and MI_COUNT=5 informs SUDAAN to use all five datasets (IMPN1, IMPN2, IMPN3, IMPN4, IMPN5) for pooling the estimates from the five logistic models. The statements NEST and WEIGHT are set for non-survey data that are nested within clinics (CLINIC). The CLASS statement is used to specify the categorical covariates and the REFLEVEL statement specifies the reference level for each categorical variable. Note with DESIGN=WR and the NEST and WEIGHT statements as listed, the variable CLINIC is modeled as a random effect.

Results

There were 335 cases with an Apgar score less than seven found in 16,833 infants in the data. The primary risk factor of interest was a three level (tubal obstruction only, ovulatory dysfunction only, and other reasons) variable of infertility diagnosis (reason for ART, V_1). The primary interest was in comparing women with ovulatory dysfunction only to women with tubal obstruction only, controlling for other covariates mentioned above. Using imputed data, all 335 cases were included in the adjusted model; however, only 268 cases and 15,430 infants could be used for the adjusted model derived from the original non-imputed data (20.0% less cases and 8.3% less infants). For our multivariable logistic model, the inspection of collinearity using Zack's SAS Macro showed that only one condition index is greater than 30, indicating no sign of multicollinearity between covariates.

The odds ratios, 95% confidence intervals (CI), and *P* values for the unadjusted and adjusted models for reason for ART are compiled in Table 1.

Comparing a diagnosis of only ovulatory dysfunction to only tubal factor, the unadjusted odds ratio (OR) using all 335 cases was 1.86 (95% CI: 1.31-2.63, P-value = 0.0005). Notice that the missing for V1 was negligible (comparing the imputed data adjusted odds ratio to the non-imputed data adjusted odds ratio) and no cases were deleted from the unadjusted analysis. Using the multiply imputed data, the adjusted odds ratio was 1.93 (95% CI: 1.31-2.84, P-value = 0.0009) and using the non-imputed data, the adjusted odds ratio was 1.73 (95% CI: 1.12-2.69, P = 0.015).

Table 1. Unadjusted odds ratio (OR) and adjusted odds ratio (aOR) for reasons for ART

Reason for ART	OR (95% CI*) P value	Imputed data aOR (95% CI*) P value	Non-Imputed data aOR (95% CI*) P value
Tubal Obstruction only	Ref	Ref	Ref
Ovulatory Dis-function only	1.86 (1.31-2.63) 0.0005	1.93 (1.31-2.84) 0.0009	1.73 (1.12-2.69) 0.015
Other reasons	1.20 (0.85-1.69) 0.297	1.35 (0.91-1.99) 0.134	1.27 (0.91-1.77) 0.152

**CI-Confidence interval*

Because there were a small number of infants with Apgar scores less than 7 (335/16,833), there was a concern that missing values of covariates would change the results of the adjusted model. This concern was addressed using the method of multiple imputation. Because the data were naturally clustered, consider the impact of such data structure in multiple imputation and modeling, which likely provides better statistical inferences than not addressing such impact on analysis. The SUDAAN HOTDECK procedure imputed missing values by incorporating covariate information in the imputation process. The merit of this approach is to use real (and hence realistic) values in imputation without strong parametric assumptions, and to provide good inferences for linear and non-linear statistics (Andridge & Little, 2010). However, this procedure has limitations, because it requires good matches of respondents to recipients based only on available covariate information and finding good matches is more likely in large clinics. Moreover, repeating the HOTDECK with the same respondent pool but randomly sorting data is an arguable imputation procedure. To determine the impact of this method on the results, we also conducted the analysis using the traditional complete observations method. In this study, the results were similar, meaning

multiple imputation may not be necessary. However, the conclusion does not exclude the possibility that results may vary across applications.

The data had a hierarchical or nested data structure with observations (infants) clustered within fertility clinics. The impact of this data structure was addressed in the multiple imputation and statistical analysis using the SUDAAN software package. The example provided could be applied to other datasets with hierarchical or nested structures where missing values of variables are a concern.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40–64. doi: 10.1111/j.1751-5823.2010.00103.x

Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. (2014). *2012 Assisted Reproductive Technology Success Rate*. Atlanta, GA: Centers for Disease Control and Prevention.

Cox, B. (1980). The weighted sequential hot deck imputation procedure. In *American Statistical Association Proceedings* Survey Research Methods Section, pp. 721–726. Alexandria, VA: American Statistical Association. Retrieved from http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1980_152.pdf

Grigorescu, V., Zhang, Y., Kissin, D., et al. (2014). Maternal characteristics and pregnancy outcomes after assisted reproductive technology (ART) by infertility diagnosis: ovulatory dysfunction (OD) versus tubal obstruction (TO). *Fertility and Sterility*, 101(4), 1019–1025. doi: 10.1016/j.fertnstert.2013.12.030

Iannacchione, V. (February, 1982). *Weighted sequential hot deck imputation macros*. Paper presented at the Seventh Annual SAS User's Group International Conference, San Francisco, CA.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2), 137–163. doi: 10.1093/oxfordjournals.pan.a004868

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, *50*(4), 1163–1170. doi: 10.2307/2533454

Rubin, D. B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. In *American Statistical Association Proceedings* Survey Research Methods Section, pp. 20–34. Alexandria, VA: American Statistical Association. Retrieved from http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1978_004.pdf

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. NY: John Wiley and Sons. doi: 10.1002/9780470316696

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. doi: 10.1080/01621459.1996.10476908

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. doi: 10.1177/096228029900800102

Sinharay, S., Stern, H., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*(4), 317–329. doi: 10.1037/1082-989x.6.4.317

Zack, M. (n.d.) SAS Macro [Computer software]. Retrieved from http://schick.tripod.com/collin.sas

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 753-774.
doi: 10.22237/jmasm/1493599200

ISSN 1538 − 9472

Selection of Statistical Software for Data Scientists and Teachers

Ceyhun Ozgur
Valparaiso University
Valparaiso, IN

Min Dou
Valparaiso University
Valparaiso, IN

Yang Li
Valparaiso University
Valparaiso, IN

Grace Rogers
Valparaiso University
Valparaiso, IN

The need for analysts with expertise in big data software is becoming more apparent in today's society. Unfortunately, the demand for these analysts far exceeds the number available. A potential way to combat this shortage is to identify the software sought by employers and to align this with the software taught by universities. This paper will examine multiple data analysis software – Excel add-ins, SPSS, SAS, Minitab, and R – and it will outline the cost, training, statistical methods/tests/uses, and specific uses within industry for each of these software. It will further explain implications for universities and students.

Keywords: Big data, Excel, R, SAS, SPSS, statistical software, data scientist

Introduction

In the age of big data, technology has transformed how business decisions are made. According the McKinsey Global Institute, "decision making will never be the same; some organizations are already making better decisions by analyzing entire datasets from customers, employees, or even sensors embedded in products" (Manyika et al., 2011, p. 5). In addition to intuition and judgment, business personnel use various software to draw conclusions from data sets and to thereby make decisions.

In measuring the popularity of many data analysis software, Muenchen (2014) noted discovering the software skills that employers are seeking would "require a time consuming content analysis of job descriptions" (para. 17). Muenchen found other ways to determine the statistical software skills that

Dr. Ozgur is a Professor of Information and Decision Sciences in the College of Business. Email him at: ceyhun.ozgur@valpo.edu.

employers seek. One method is to examine which software they currently use. Muenchen cited a survey conducted by Rexer Analytics about the relative popularity of various data analysis software in 2010. The results are pictured in Figure 1. Data miners use R, SAS, and SPSS the most. It can be inferred these are the software skills that the greatest proportion of employers will continue to look for in their potential employees. However, this method only examines the software that employers might seek if they are hiring, so it does not accurately measure the software that they currently look for in their current employees.

Another method used by Muenchen (2014) was to study software skills employers currently seek as they try to fill positions. A rough sketch of statistical software capabilities sought by employers was put together by perusing the job advertising site Indeed.com, a search site the comprises major job boards – Monster, Careerbuilder, Hotjobs, Craigslist – as well as many newspapers, associations, and company websites. The results are summarized in Figure 2.

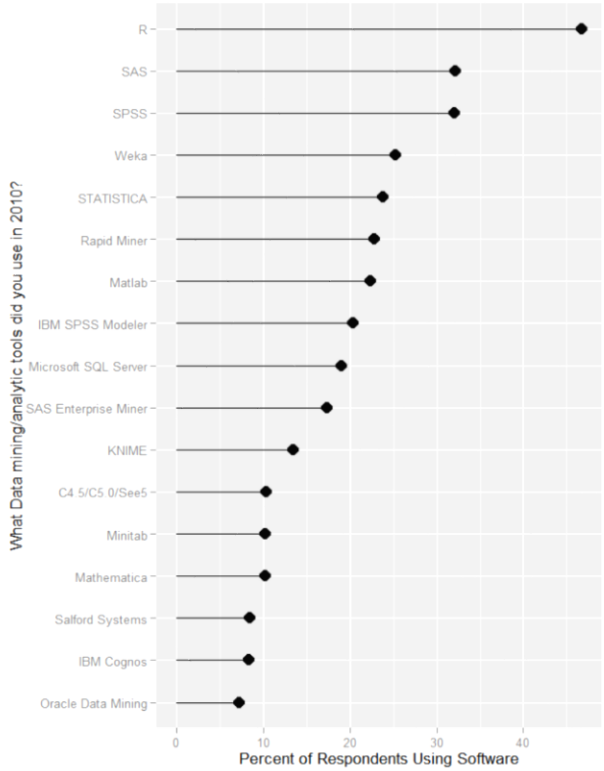


Figure 1. 2010 Rever Analytics survey results of analytic tools (Muenchen, 2014)

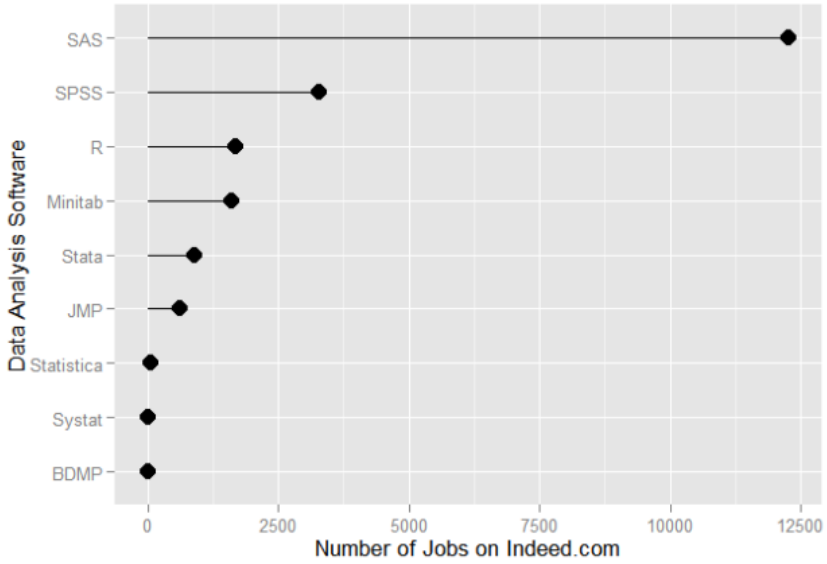


Figure 2. Jobs requiring various software (Muenchen, 2014)

As seen, in contrast to R's greater usage in companies over SAS, illustrated in Figure 1, jobs requiring SAS led to more positions than any other data analysis software. For employers, SPSS and R skills finished in second and third place. This second estimation method of Muenchen (2014) measured the software skill deficits in the job market. It seemed the demand for people with SAS skills outweighs the number of individuals with this capability. One reason for this disconnect could be that college and university faculty are not teaching SAS skills in proportion to the demand for these skills (Lofland & Ottesen, 2013).

To assess this potential disconnect, a non-random survey was conducted with faculty from eighteen departments, which included small and large, state and private, undergraduate and graduate, and East and West, with the results compiled in Table 1. As expected, there was a discrepancy between the software taught and the software sought. SAS led in job openings, but data analysis software taught at those universities did not reflect it. Only a few departments had faculty who taught SAS more than R or SPSS.

The faculty at some departments did not teach any software at all. For example, at Valparaiso University, faculty in the Information and Decision Sciences Department did not teach statistical software, although in certain courses the faculty utilized SPSS, SAS, and R. Excel was the most applicable software used.

Table 1. Results from a survey of statistical software packages taught (Compiled by Kleckner, 2014)

Department	Software Taught at Grad Level	Software Taught at Undergrad Level
		Large, Midwestern, State Universities
Actuarial Science	SAS, Excel, Mathematica	SAS
Mathematics	None	none
Marketing	SAS, SPSS, JMP	N/A
Marketing	SPSS, Excel*	SPSS, Excel*
		Large, Southeastern, State Universities
Statistics	SAS, R, SAS Enterprise Miner	SAS, R, JMP
Engineering	Excel, JMP, Matlab, Mathematica, Mathcad	SAS, Excel, JMP, Matlab, Mathematica, Maple, Mathcad
Economics	N/A	SAS, R, ForecastX, GRETL
Economics	No Graduate Program in Economics	SPSS, Excel, Stata
Information Systems & Decision Sciences	SAS, SPSS, Excel, Megastat, JMP, SAP, Minitab, Matlab, Stata, Mathematica*	SAS, SPSS, Excel, Megastat, JMP, SAP, Minitab, Matlab, Stata, Mathematica*
		Medium, Northeastern, Private Universities
Statistics	SAS, R, Excel, Minitab, JMP, Matlab, Python	N/A
Mathematics	SAS, R, JMP, Matlab, DataDesk, ActivStats*	SAS, R, JMP, Matlab, DataDesk, ActivStats*
		Medium, Southeastern, Private University
Biostatistics	SAS, SPSS, Minitab, Mathematica, Fortran, StatExact, Spatial Stat, C, C++	No Undergraduate Program in Biostatistics
		Small, Midwestern, Private Universities
Mathematics & Computer Science	N/A	SAS, Excel
Mathematics	No Graduate School	SPSS, Excel, Minitab, Mathematica
Statistics	No Graduate School	R
Economics	No Graduate School	Minitab, GRETL
		Small, Southern, Private Universities
Computational and Applied Mathematics	Matlab, C, C++	Matlab, C, C++
Statistics	SAS, SPSS, R, Excel, JMP, Matlab, Mathematica, Stata	JMP, Stata

Note: * These schools did not specify whether the software listed were for graduate or undergraduate students, so we assumed both

This survey was not random, and therefore they cannot be generalized throughout the United States. However, within the sample, there was a trend seen in quantitative, engineering, and business departments, where the use of statistical packages were not aligned to the skills required by employers.

Paying attention only to job availability, it seems that many schools need to reconsider their software choice in favor of implementing SAS. Nevertheless, there are many factors to consider other than the popularity within the job market. Faculty must also consider the cost and time effectiveness of incorporating each software into their curriculum. Further, faculty in specific departments within the school should consider which software best fits their area of study.

Purpose of the Study

The purpose of this study is to gather and condense the necessary information for teaching statistical software. It will assist faculty in their software choices, and it will help their counterparts in business decide which software is best to bring their workforce to the next level of capability. This has increased importance as big data analysis becomes a necessity in business, as Manyika et al. (2011) noted.

The impact of developing a superior capacity to take advantage of big data will confer enhanced competitive advantage over the long term and is therefore well worth the investment to create this capability. But the converse is also true. In a big data world, a competitor that fails to sufficiently develop its capabilities will be left behind. Big data can no longer be ignored, as noted by the successes of companies where it is invoked as compared to less-modern competitors (Manyika et al., 2011).

Computer software can be written to flexibly support statistical practice (Buchan, 2000). Hence, the focus of this study is on SAS, SPSS, and R software, because both methods in Muenchen's (2014) study indicated they are the three most competitively sought software in business.

Minitab for Teaching Purposes

Minitab's Quality Trainer teaches users how to analyze data online. This multimedia course includes animated lessons that bring statistical concepts to life, and interactive quizzes that give real-time feedback. Hands-on exercises walk the user through applying statistics with Minitab Statistical Software, so knowledge may be put to use immediately.

Quality Trainer contains nine chapters with 141 interactive lessons that can be repeated. It covers statistics needed to analyze quality improvement data, including Basic Statistics, Control Charts, Process Capability, ANOVA, DOE, and more. Easily implementation of projects using a comprehensive collection of more than 100 tools specifically designed for each task. These built-in templates promote greater speed and accuracy.

Below is a list detailing Minitab guide use of statistical and other tools to manage projects:

- **Value Stream Mapping:** Establish the flow of materials and information through your organization. Streamline processes to add value that meets customer expectations.
- Fishbone Diagram: Identify every relevant element of your process and refine the scope of complex projects.
- On-Demand Coaches: Receive the expert guidance you need to complete every step of your project. Add your own instructions or information to any Coach.
- Process Mapping: Construct high-level or detailed flow charts that help you understand and communicate all the activities in a process. Assign variables to each shape and then share them with other tools.
- FMEA (Failure Modes and Effects Analysis): Identify the potential causes for a product or process failure, anticipate the resulting effects, and prioritize the actions needed to mitigate them.
- Pugh Matrix: Compare product design proposals and improvement strategies and determine which ones best fulfill your customer requirements and organizational goals.
- Capture Analysis: Identify and record the important and relevant sections of your Minitab analyses.
- Financial Analysis: Estimate your project savings and the timeframe for realizing them.
- Project Risk Assessment: Evaluate whether a potential project can be successfully completed on time.
- Stakeholder Analysis: Summarize the impact your stakeholders have on your project so you can more effectively leverage their support and address their concerns.
- 5S Audit: Evaluate process conditions relative to 5S best practices and track the ongoing implementation of 5S improvements and controls.

- SIPOC (Supplier-Input-Process-Output-Customer) Analysis: Identify every relevant element of your process and refine the scope of complex projects.
- C&E (Cause and Effects) Matrix: Save time determining what *X* variables to address by comparing and evaluating their potential to impact your goal.
- Y Metrics Chart: Evaluate the progress of your project over time in relation to its baseline and goal.
- Insert Team Members: Easily add team members to your project from your e-mail address book or other file.

Excel Add-Ins

Add-ins are programs that add optional features and commands. With regard to Microsoft Excel, there are add-ins for a multitude of purposes: data analysis, presentation, investment, business, personal, utilities, and productivity tools, and organization. Within data analysis are the Analysis Toolpak, Solver, MegaStat, and PHStat. Both MegaStat and PHStat access codes come with a textbook. However, if an access code isn't available for PHStat, the MegaStat add-ins are available separately from McGraw-Hill (http://highered.mheducation.com/sites/0077425995/information_center_view0/index.html) and Pearson (https://wps.aw.com/phstat/), respectively. Although the Analysis Toolpak and Solver are free add-ins, MegaStat is not.

MegaStat Training

With the current focus on STEM (science, technology, engineering, and mathematics), students and workers may already be familiar with Microsoft Excel or similar spreadsheet software. Building on this familiarity, Burdeane (O. Burdeane, personal communication, January 29, 2014) explained, "Since MegaStat looks and works like Excel, almost anyone could use it to generate some output with just a few minutes of training". MegaStat has dialog and input boxes, buttons, and checkboxes that work largely the same as those in Excel. Therefore, the 53-page tutorial PDF – complete with a step-by-step process to using each test that MegaStat performs, and pictures at every step – will likely provide sufficient guidance to effectively use this software.

Statistical Methods/Tests/Uses

MegaStat can perform a multitude of statistical operations: descriptive statistics, frequency distributions, probability, confidence intervals and sample size, hypothesis tests, ANOVA, regression, time series/forecasting, chi-square, nine nonparametric tests, quality control process charts, and generate random numbers (McGraw-Hill Education, 2014). SPSS and SAS, for example, have more advanced options, "especially in the area of multivariate statistics" (O. Burdeane, personal communication, January 29, 2014). However, "MegaStat can handle most things encountered by non-PhD statisticians" (O. Burdeane, personal communication, January 29, 2014).

The major caveat for this inexpensive and easy-to-use software is its size capability. For example, Burdeane (O. Burdeane, personal communication, January 29, 2014) experimented with the number of data points that MegaStat can handle, and noted, "I did find a file with 10 columns and 152630 rows. That is over 1.5 million data points and MegaStat did a descriptive statistics analysis on it in about 10 seconds." Although the capability to analyze a million and a half data points sounds quick, this capability may not meet the demand of large companies, because "Wal-Mart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data," and "Facebook handles more than 250 million photo uploads and the interactions of 800 million active users with more than 900 million objects (pages, groups, etc. – each day" (Troester, 2012, p. 1). Extracting this data and making use of it using MegaStat is not feasible. Other restrictions of MegaStat include its limitation to twelve independent variables in multiple regression and restrictions on variables and table size (O. Burdeane, personal communication, January 29, 2014).

Burdeane (personal communication, January 29, 2014) opined

> I would guess that most use of MegaStat in companies is by people who are not professional statisticians. I think people with formal training in statistics beyond an introductory course would have experience with one of the big packages (SAS, SPSS, Minitab) and would tend to stick with that software even if it was overkill for many analyses.

Burdeane also suggested many analyses do not require major packages, like SAS, SPSS, and R, but statisticians stick to them because they are comfortable.

However, personnel in industry still use Excel. For example, a global appliance manufacturer uses Excel "for extensive 'What If' analysis around budgeting" and to forecast (J. Ward, personal communication, January 20, 2014).

Other Excel Add-Ins

- Analyse (www.analyse-it.com) Standard Edition
- XLStat (www.xlstat.com) from Addinsoft's website.
- NumXL (www.spiderfinancial.com/products/numxl)
- Quantum XL (www.sigmazone.com)

SPSS

SPSS, originally termed Statistical Package for the Social Sciences, was released in 1968 as a software designed for the social sciences. A series of companies subsequently acquired SPSS, ending with International Business Machines (IBM), the current owner, during which time the product's user base was expanded. Therefore, its former acronym was replaced with Statistical Product and Service Solutions to reflect the greater diversity of its clients. Along with Minitab, it is one of the leading statistical packages used in the social and behavioral sciences.

Cost

Consumers can buy SPSS software packages separately by choosing a particular product that they think will satisfy their need; however, SPSS offers bundles that cost much less than paying for the programs independently. SPSS offers three of these bundles: standard, professional, and premium.

Within each of these bundles, SPSS gives four options: an authorized user license, authorized user initial fixed term license, concurrent user license, and concurrent user initial fixed term license. Thus, when customers decide they want to purchase SPSS, they have to make two decisions: user license versus initial fixed term license, and authorized user versus concurrent user. User licenses never expire, while initial fixed term licenses last for twelve months. An authorized user is a single licensee who buys the right to use the program; a concurrent user is the right for a single person to use the program at a given time, but it does not distinguish who this person has to be.

SPSS also offers student packages for college attendees. Students can purchase the single user initial fixed term license "SPSS GradPack" software from their college or university, or they can buy it from SPSS's official

distributers, like Creation Engine, On the Hub, StudentDiscounts.com, Studica, or ThinkEDU (IBM, n.d.c). For example, on the Creation Engine website, students can buy the SPSS Statistics Premium GradPack (IBM, n.d.a).

Training

Crossman (n.d.) addressed the difficulty of using SPSS for the first time:

> SPSS provides a user interface that makes it very easy and intuitive for all levels of users. Menus and dialogue boxes make it possible to perform analyses without having to write command syntax, like in other programs. It is also simple and easy to enter and edit data directly into the programs. (SPSS section, para. 1)

Although SPSS does look similar to typical spreadsheet applications like Excel, and its ease of use is very comparable to Excel as well, the cells cannot be manipulated in spreadsheet fashion.

Statistical Methods, Tests, Uses

"SPSS was designed specifically for statistical processing of large amount of data at an enterprise level," while spreadsheets are broadly applicable to many different tasks outside of statistical computing (Robbins, 2012, para. 3). An advantage of this specialized design is that SPSS "keeps calculated statistics and graphs separate from the raw data but still easily accessible" (Robbins, 2012, para. 3). SPSS software furthermore has a much more convenient platform for performing statistical tests. For instance, performing a one-sample *t*-test in Excel (without a plug-in) requires some independent calculations by the user, whereas with SPSS, the user only needs to "select a variable and supply the value to compare with [the] sample" and click "Ok" (Robbins, 2012, para. 4). Another advantage of SPSS is that it links numerically coded data to its original meaning (Robbins, 2012). With most data being electronically stored in numerical fashion, this feature of SPSS is highly valuable.

SPSS's standard bundle includes its statistics base, advanced statistics, bootstrapping, custom tables, and regression capabilities. Purchasing the professional bundle further supplies the consumer with the categories, data preparation, decision trees, forecasting, and missing values features. The most comprehensive bundle, premium, provides the user with the complex samples, conjoin, direct marketing, exact tests, neural networks, amos, sample power, and

visualization designer, in addition to all of the packages from the professional bundle (IBM, n.d.b). SPSS is also useful for generating plots of distributions and trends, charts, and tabulated reports.

Specific Uses in Industry

On its website, prospective SPSS clients can read about applications in fields like automotive, banking, chemical and petroleum, computer services, consumer products, education, electronics, and energy and utilities. They can also access a list of SPSS's clients. Below are specific examples of SPSS at work within business.

- Infinity Insurance uses SPSS's predictive analytics feature to detect fraudulent claims (IBM, n.d.e).
- "By mining alumni and stakeholder records, social media and other unstructured data-sets with text analytics software, [Michigan State University] gains insights into the engagement, sentiments and behavior of current and potential donors," which enables smarter fundraising (IBM, n.d.d).
- The Guardia Civil, Spain's very first national law enforcement agency, has investigated crimes and psychology using SPSS (IBM, n.d.d).
- One distinguished hospital uses SPSS to forecast payment behavior. It tries "to better identify patients who are most likely to pay their hospital bills" by what it calls "predict[ing] patient payment potential" (IBM, n.d.d).

SAS

SAS (Statistical Analysis System) is a commercial statistical package that was developed during the 1960s at North Carolina State University as part of an agricultural research project. Its usage has grown considerably. Ninety-one of the top one hundred companies on the 2013 Fortune Global 500 list use the software (SAS Institute, n.d.a). SAS does not run on Mac computers very easily. One way to run the software on a Mac computer is through parallels, where users buy and run the Windows interface as well.

Cost

An individual license of the Analytics Pro version is available on an annual basis, with a price reduction for subsequent years. With a few more features than the Analytics Pro system, the Visual Data Discovery package is more expensive, although renewals are available with a price break. It is important to consider the added costs if a user wishes to perform data analysis for the benefit of some other party. A different license must be obtained by consulting a SAS representative (SAS Institute, n.d.c).

One of these alternative licenses is a server-based license. These licenses certainly save schools and businesses money by allowing their affiliates each to access the software through a web-based connection or a network. SAS fills these requests on a case-by-case basis, so interested customers should speak to SAS directly to get a quote (SAS Institute, n.d.c).

On top of these two versions, SAS has created an OnDemand edition, which is available at no cost to degree granting institutions. Professors can set up an account online, and they and their students can access the software anywhere with an Internet access. Although this free software "has been reported to be slow at times," it definitely provides a great opportunity for schools to teach students the basics of SAS programming (Lofland & Ottesen, 2013, p. 3).

In addition to the software license, there is also considerable cost time in the form of installation. Lofland and Ottesen (2013) explained that "SAS can be difficult for users to obtain and the initial installation is sometimes tricky […] long and difficult" (p. 3). However, SAS does not require users to install additional packages.

Training

Crossman (n.d.) claimed, "SAS is a great program for the intermediate and advanced user because it is very powerful, can be used with extremely large data sets, and can perform complex and advanced analyses" (SAS section, para. 1). SAS requires more training than Excel and SPSS, because it largely runs on programming syntax rather than point-click menus that other software boast.

The amount of training necessary for individuals to properly use SAS depends on many factors, including the trainee's background and the type of analysis she will need to perform. In terms of background, prospective SAS programmers with prior programming experience will have a much easier time. SAS syntax resembles that of other programming languages, so experience with one language often helps learn another. For instance, SAS is similar to Java in that

both contain data values, function calls, identifying key words at the beginning of each line, and semicolons at the end of each line (Boudreaux, 2003, p. 1). However, even if the syntax of SAS and a previously learned language are completely different, experience with coding is extremely helpful because the art of programming is a different kind of thinking. The training required also depends on the type of analysis that the trainee must carry out. If the trainee only needs to run the same type of test repeatedly, then she may only need training in a specific aspect of SAS programming; however, if the trainee will need to develop a process based on each new task, then she will need more sound understanding of the software.

Fortunately, experts have written copious texts about how to use SAS and SAS has a strong user support system; even if users do not have complete understanding of the software, they can run it. Although there exists no easy way to calculate the number of books written about SAS, Muenchen estimates it by searching for books published with SAS in their title and found that close to 500 were published between 2001 and 2011 (as cited in Lofland & Ottesen, 2013). Regarding user support, Lofland and Ottesen (2013) observed

> SAS has extensive online documentation, expert technical support, professional training courses, many excellent books in press, and a tight knit user group and web based community. Problems can be addressed to SAS directly via tech support who replies very quickly and will work with the user to solve the problem. (p. 3)

They designated the user support service of SAS as one of its main specialties. Therefore, even though SAS requires some programming skill, the strength of SAS's support system makes it more manageable for less advanced users.

Statistical Methods, Tests, Uses

SAS's Analytics Pro bundle comes with three of the most popular SAS products: Base SAS, SAS/STAT, and SAS/GRAPH. The Visual Data Discovery collection includes SAS Enterprise Guide (SAS's only point-click interface) and JMP software to make discovery and exploratory analysis easier.

With either of these toolsets, programmers can perform a number of statistical tests. The Institute for Digital Research and Education website outlines a multitude of statistical tests and their corresponding SAS code. The list includes thirty-two tests that come from statistical categories such as regression, factor

analysis, discriminant analysis, ANOVA, non-parametric tests, and correlation (University of California, Los Angeles [UCLA]: Statistical Consulting Group, n.d.). The full list can be seen in the Appendix below.

SAS can perform many more statistical tests than just these, though. It also functions well with forecasting, time series analysis, and many other advanced statistical techniques. In fact, SAS has created specialized programs for these methods. The SAS website's "Products & Solutions" (http://support.sas.com/software/) has a list of these programs.

Also on this page, SAS has additional packages to access that are industry-specific. For example, there is an SAS Drug Development package that "enables the efficient development, execution and management of analysis and reporting activities for clinical research," (SAS, n.d.b) an SAS Fraud Management package that "delivers a full-service enterprise-wide fraud management system that offers real-time scoring of accounts by looking at all card transactions—including purchases, payments and nonmonetary transactions," (SAS, n.d.b) and an SAS Risk Management for Insurance package that "implements the Solvency II standard model approach for calculating risk-based capital with [its] comprehensive solution for preforming risk analysis and risk-based capital calculations" (SAS Institute, n.d.b). In addition to these specialized packages for health-care, banking, and insurance, SAS has formulated software with built-in functions for other areas like law enforcement, communications, retail, casinos, utilities, and sports, among others.

SAS's advantageous functions extend beyond just carrying out statistics, though. It has superior qualities for both before the statistical analysis and after. Prior to the actual statistics, it facilitates the reading in and managing of disorganized data. Real life data is rarely clean and analysis-ready. SAS can interpret messy data sets, convert them to a clean form, and manipulate them in ways that other software cannot (Lofland & Ottesen, 2013, p. 3-4). After the user performs the statistics, SAS has impressive graphics and report-writing features that will help disseminate the findings in clear and appealing ways. But, these aesthetic products come with a caveat according to Lofland and Ottesen (2013), who explained, "SAS provides many useful procedures for creating detailed and polished reports," however, "some of the more detailed reporting procedures […] have a learning curve that takes place before being able to use them correctly" (p. 3-4).

Specific Uses in Industry

SAS has built-in, functional packages for many specific industries, including health-care, banking, insurance, law enforcement, communications, retail, casinos, utilities, sports, and more. To follow are a couple of real-life uses of SAS within some of these industries.

- A leading medical device company utilizes SAS "for clinical study data analysis" (K. Kleckner, personal communication, February 1, 2014). This same company furthermore uses the software "for setting sample sizes for pre-clinical studies and human clinical studies; [and] for setting controls on manufacturing operations" (K. Kleckner, personal communication, February 1, 2014).
- A global appliance manufacturer uses SAS for quality control by performing predictive analyses of product defects (J. Ward, personal communication, January 20, 2014).

R

R is a free, open-source statistical software. Colleagues at the University of Auckland in New Zealand, Robert Gentleman and Ross Ihaka, created the software in 1993 because they mutually saw a need for a better software environment for their classes. R has more than two million users according to an R Community website (Revolution Analytics, n.d.a).

Cost

R is free and is downloadable from the Internet, with no subscription fees, user limits, or license managers. However, this presents a danger. As open source software, R could be a security concern for large companies, because the software can be freely used, changed, and shared by anyone.

Like SAS, R can be expensive in a form other than monetary. Although the base for R is very easy to install, users must download packages to perform specific analyses, which can be very time-consuming (Lofland & Ottesen, 2013, p. 3-4). For example, currently there are 5,508 available packages, and this number grows weekly if not daily (Comprehensive R Archive Network [CRAN], n.d.). This provides many options, but searching through the assemblage of choices can be difficult and time-consuming.

Training

The training necessary for effectively using R depends on the previous computing experience of the trainee. Computing experience is helpful because data analysis in R requires writing functions and scripts, not just pointing and clicking. In many ways, though, R is comparable to other programming languages. For instance, similar to many other languages, it is a command line interface. Additionally, its source code is similar to that of C and Fortran, and it supports matrix arithmetic and data structures like APL and MATLAB. Having used any of these in the past could lessen the training time necessary to learn R. As stated with SAS above, though, having any programming experience at all will often speed up the learning process for trainees since programming problems are a completely different type of puzzle.

Sources report varied answers when identifying the training necessary to successfully utilize R. Some believe that R does not necessitate much knowledge of computer programming after all. For example, Pregibon claimed R "allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems (Vance, 2009, para. 4). Vance (2009) also noted, "R has quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use" (para. 3). R is not as daunting as other languages, having very natural and expressive syntax for data analysis. In R language, "anova(object_1, object_2)" produces an ANOVA table; "coef(object)" extracts the regression coefficient; and "plot(object)" produces plots showing residuals, fitted values, and other diagnostics (R Core Team, 2014). Still, R does require the use of objects, operators, and functions before applying these intuitive commands. Fortunately – as stated earlier – many packages are available for download and use off the Internet, so users do not necessarily have to know the code or write it. This is another reason why some say that R does not require much programming knowledge.

However, because of errors in some of these packages and lack of user support for R, others believe that advanced training investment is necessary in order to use the software. Lofland and Ottesen (2013) stated, "[R] users rely on what others put out there about the software. [….] Packages are not written by the R Development Core-Team; therefore, they are not well polished and some could have questionable validity. It is also difficult to direct an issue to a particular person or support system" (p. 3). Although R may be useable without much coding experience, when a problem arises, the lack of programming knowledge will become evident and costly due to a dearth of documentation and technical support for resolving the issue. In other words, people without sufficient

knowledge of the R programming language can implement the syntax in their own use, but they do not necessarily have solid understanding of what the code actually says. This lack of R coding knowledge makes debugging difficult if not impossible, and it could lead to erroneous results with severe decision-making consequences.

Lofland and Ottesen (2013) also explained that report writing in R is difficult. They claim that the extensive programming required to code a report in R is quite a time investment, as "R does not have a defined way of producing reports" (p. 3).

Statistical Methods, Tests, Uses

R is a comprehensive statistical analysis toolkit. It can perform any statistical analysis desired, but users must either write the code or access the code from someone who has already written it. As stated on its website, people have already designed many standard data analysis tools "from accessing data in various formats, to data manipulation (transforms, merges, aggregations, etc.), to traditional and modern statistical models (regression, ANOVA, GLM, tree models, etc.)" (Revolution Analytics, n.d.b). Programmers have designed many more packages than just these, including packages for Bayesian statistics, time series analysis, simulation based analysis, spatial statistics, survival analysis, and many, many more (CRAN, 2014). A complete list of packages already designed for R can be found on the R packages website (http://cran.us.r-project.org/web/packages/).

The key feature of R that differentiates it from other statistical software is its acceptance of customization. The aforementioned software have "data-in-data-out black-box procedures" (Revolution Analytics, n.d.b). Developers have written the code for a certain function, such as performing decomposition for a time-series model, and users have never seen this built-in code that runs in the background. A "decomp" command, or something of the sort, is all that is needed, and the statistical package will perform the decomposition for them. For example multiplicative decomposition is the forecasted value (F) = Trend × Seasonal × Cyclical × Irregular. However, R is an interactive language. It requires users to write the code (for the decomposition, or whatever procedure desired) or to paste the code in from someone who already wrote it. Because the function's code is visible in their command box, users can manipulate the commands however they see fit. Thus, R enables experimentation and exploration by allowing users to improve the software's code or to write variations for specific tasks. They can

even mix-and-match models for better results. With the pre-packaged functions in the other statistical software, this is not as easy.

R is known for generating appealing charts and tables. The custom charting capabilities of R create "stunning infographics seen" (Revolution Analytics, n.d.a). However, it cannot manage messy data as easily as other available statistical software. Lofland and Ottesen (2013) warned, "The design of R was focused around statistical computing and graphics, so data management tends to be time consuming and not as clean as SAS. […] Students who have used solely R have an unrealistic expectation of the state of the data they receive" (p. 3). But, once the data is organized, R is a valuable data analysis performer and graphics creator.

Specific Uses in Industry

The usage of R is diverse in business. Some examples follow.

- Google "taps R for help understanding trends in ad pricing and for illuminating patterns in the search data it collects" (Vance, 2009, para. 24).
- Pfizer has engineered its own custom packages in R, which allows scientists to manipulate their own data during nonclinical drug studies instead of hiring a statistician to do the work for them (Vance, 2009).
- A financial services company utilizes dozens of R packages to perform derivatives analysis (Vance, 2009).

Conclusion

Excel add-ins are well-suited to small companies and small projects because of their availability and low cost, while SPSS, SAS, and R work well for large projects and large businesses because of their ability to handle large sums of data efficiently. As discovered at the beginning of the paper, Excel's MegaStat option can execute many important statistical procedures that people trying to interpret smaller data sets can utilize for low financial cost and training cost. However, as stated, MegaStat can only manage a certain amount of data. Therefore, larger data sets require a higher-powered software, like SPSS, SAS, or R. Differentiating between which of these software best fits the analysis of these larger data sets depends on a number of factors, and each statistical package has its own strengths and weaknesses. Hence, the purpose of this study was to investigate their features.

Finding the suitable software is important, because companies that employ the most efficient data analysis software will compete better against competition

by effectively accessing and using their stockpiles of data to make better decisions. Faculty at colleges and universities could improve job placement by preparing students in the specific software that hiring companies use. It is difficult for new data analysts to see the forest for the trees when choosing a statistical programming language (DataCamp Team, 2014).

Students can add a software taught category to their list of traits sought in higher education in order to prepare themselves for job placement. One of the most important decisions that future students make is selecting a major. Often, a student's desired major can influence the selection set. However, other decisions are growing in importance too. In terms of finding a job, employers are increasingly seeking out recent graduates that have experience with big data software, like SPSS, SAS, and R. Therefore, it is becoming more important for students to seek out a university that will prepare them with knowledge of pertinent software, which will increase their likelihood of finding a satisfying job. Obviously, careers in big data will be abundant, so prepared students will have little trouble finding a job in that area. Nevertheless, students trained on high demand software will have more and better options for job placement.

References

Boudreaux, D. (2003, March-April). *Java syntax for SAS programmers*. Paper presented to the SAS Users Group International, Seattle, WA.

Buchan, I. E. (2000). *The development of a statistical computer software resource for medical research* (Doctoral dissertation). University of Liverpool, Liverpool, England.

Comprehensive R Archive Network. (n.d.). *Contributed packages*. Retrieved from http://cran.us.r-project.org/web/packages/

Crossman, A. (n.d.). *Analyzing quantitative data: Statistical software programs for use with quantitative data* [Web log post]. Retrieved from http://sociology.about.com/od/Research-Tools/a/Computer-programs-quantitative-data.htm

DataCamp Team. (2014, June 3). What is the best statistical programming language? [Web log post]. Retrieved from https://www.datacamp.com/community/tutorials/statistical-language-wars-the-infograph#gs.eO=sntU

International Business Machines. (n.d.a). *IBM SPSS Statistics Premium GradPack 22*. Retrieved from

http://www.creationengine.com/html/p.lasso?l=SPSS%20Statistics%20Premium%20GradPack&p=18830

International Business Machines. (n.d.b). *SPSS Statistics*. Retrieved from http://www-01.ibm.com/software/analytics/spss/products/statistics/buy-now.html

International Business Machines. (n.d.c). *SPSS Statistics GradPack*. Retrieved from http://www-03.ibm.com/software/products/en/spss-stats-gradpack/

International Business Machines. (n.d.d). *Success stories for SPSS*. Retrieved from http://www-01.ibm.com/software/success/cssdb.nsf/topstoriesFM?OpenForm&Site=spss&cty=en_us

International Business Machines. (n.d.e). *Why SPSS software?* Retrieved from http://www-01.ibm.com/software/analytics/spss/

Lofland, C. L., & Ottesen, R. (2013, April-May). *The SAS versus R debate in industry and academia*. Paper presented to the SAS Global Forum 2013, San Francisco, CA.

Manyika, J., Chi, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

McGraw-Hill Education. (2014). *MegaStat*. Retrieved from http://glencoe.mcgraw-hill.com/sites/0010126585/student_view0/megastat.html

Muenchen, R. A. (2014). *The popularity of data analysis software*. Retrieved from http://r4stats.com/articles/popularity/

R Core Team. (2014). *An introduction to R*. Retrieved from http://cran.r-project.org/doc/manuals/R-intro.html#Statistical-models-in-R

Revolution Analytics. (n.d.a). *What is R?* http://www.inside-r.org/what-is-r

Revolution Analytics. (n.d.b). *Why use R?* http://www.inside-r.org/why-use-r

Robbins, S. (2012, June 7). *How does SPSS differ from a typical spreadsheet application*. Retrieved from https://publish.illinois.edu/commonsknowledge/2012/06/07/how-does-spss-differ-from-a-typical-spreadsheet-application/

SAS Institute. (n.d.a). *About SAS*. http://www.sas.com/en_us/company-information.html

SAS Institute. (n.d.b). *Industry solutions*. Retrieved from http://www.sas.com/en_us/industry.html

SAS Institute. (n.d.c). *Pricing and licensing information*. Retrieved from https://www.sas.com/order/product.jsp?code=PERSANLBNDL

Troester, M. (2012). *Big data meets big data analytics: Three key technologies for extracting real-time business value from the big data that threatens to overwhelm traditional computing architectures*. Retrieved from http://www.sas.com/resources/whitepaper/wp_46345.pdf

University of California, Los Angeles: Statistical Consulting Group. (n.d.). *What statistical analysis should I use? Statistical analyses using SAS*. Retrieved from http://www.ats.ucla.edu/stat/sas/whatstat/whatstat.htm

Vance, A. (2009, January 6). Data analysts captivated by R's power. *The New York Times*. Retrieved from http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all&_r=0

Appendix A: List of Tests That SAS Can Perform

- One sample *t*-test
- One sample median test
- Binomial test
- Chi-square goodness of fit
- Two independent samples *t*-test
- Wilcoxon-Mann-Whitney test
- Chi-square test
- Fisher's exact test
- Kruskal-Wallis test
- Paired *t*-test
- Wilcoxon signed rank sum test
- McNemar test
- One-way repeated measures ANOVA
- Repeated measures logistic regression
- Factorial ANOVA
- Friedman test
- Ordered logistic regression
- Factorial logistic regression
- Correlation
- Simple linear regression
- Non-parametric correlation
- Simple logistic regression
- Multiple regression
- Analysis of covariance
- Multiple logistic regression
- Discriminant analysis
- One-way MANOVA
- Multivariate multiple regression
- Canonical correlation
- Factor analysis

(UCLA: Statistical Consulting Group, n.d.)

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 775.

ISSN 1538 – 9472

Book Reviews

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 776-777.
doi: 10.22237/jmasm/1493599260

ISSN 1538 − 9472

Book Review: Multivariate Statistical Methods, A Primer

C. R. Rao
University at Buffalo
Buffalo, NY

Multivariate Statistical Methods, A Primer*, *4th Ed. Bryan F. J. Manly and Jorge A. Navarro Alberto. NY: Chapman & Hall / CRC Press. 2016. 264 p. ISBN 10: 1498728960 / ISBN 13: 978-1498728966

The purpose of the book is to introduce multivariate statistical methods to non-mathematicians. It is assumed that readers have a working knowledge of elementary statistics, including tests of significance using normal, *t*, Chi-squared and *F* distributions, analysis of variance and linear regression. The authors made an excellent effort by presenting multivariate data of different kinds, such as body measurements, made on two or more kinds of individuals within each group and raising questions such as how different the measurements are within groups and how different they are between different kinds of individuals. With one measurement, differences between groups is examined by comparing individual mean values and variances within groups. With p measurements, p mean values are needed, and $p\,(p-1)$ variances and covariances for comparison. Appropriate multivariate methods for this purpose have been demonstrated. In addition, there is the problem of grouping given populations by similarity of measurements which needs a measure of distance between populations based on observed data.

The authors give a good account of different methods available for these purposes. Some of the measures of similarity such as Penrose and Mahalanobis distances are mentioned for possible use. Penrose distance does not take into account correlations between measurements and may not be appropriate in all practical applications. Mahalanobis distance will be appropriate for correlated variables when the measurements are nearly normally distributed. Some discussion on the choice of distance measure to be used will be helpful to

Prof. Rao, Sc.D. (Cantab), FRS, is the Eberly Professor Emeritus of Statistics and the Director of Center for Multivariate Analysis at Penn State, and Research Professor in Biostatistics at the University at Buffalo. Email him at crr1@psu.edu.

practical workers. The authors describe all available statistical methods for these purposes in terms of principal components, factor analysis, discriminant functions and canonical correlation analysis.

Multivariate analysis was developed during the 1940s. The Anthropology Department at Cambridge University, UK sent an expedition to Jebel Moya in Africa to dig ancient graves and bring the skeletons back for study. Their purpose was to analyze multiple measurements to find their relationship with skeletal material available in nearby areas. This is a multivariate problem for which no solution was available at that time.

Professor Trevor, a member of the Anthropology faculty, heard about the work of Prof. Prasanta Chandra Mahalanobis and the distance from point *P* and distribution *D* named after him. In July, 1946 he sent a telegram to Prof. Mahalanobis to send someone to Cambridge to analyze the skeleton measurements. At that time, I was working in the Indian Statistical Institute as a research scholar under the direction of Prof. Mahalanobis, and I had published some papers on multivariate analysis. Prof. Mahalanobis deputed me to go to Cambridge and analyze their data. I travelled to Cambridge that month, and for the following two years worked in Cambridge's Anthropology Department as a paid visiting scholar. The result was the development of the necessary tools to analyze their multivariate data, and were published in 1954 by Cambridge University Press as a book, *Ancient Inhabitants of Jebel Moya*, under the joint authorship of myself along with two anthropologists, Trevor and Mukherji.

Subsequently, I was asked by the President of Royal Statistical Society (RSS) to present my research work on the Jebel Moya data at a meeting of the Society, which I did in October, 1948. That material was later published in two research papers, which constitute the corpus of multivariate analysis as practiced today. One is Rao, C. R. (1948), Utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society*, *10*, 159-203. The other is Rao, C. R. (1948), Tests of significance in multivariate analysis, *Biometrika*, *35*(1-2), 58-79. This material also provided the basis for my book Rao, C. R. (1952), *Advanced Statistical Methods in Biometric Research*. NY: Wiley.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 778.

ISSN 1538 – 9472

Letters to the Editor

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 779-782.
doi: 10.22237/jmasm/1493599320

ISSN 1538 − 9472

Errors in a Program for Approximating Confidence Intervals

Andrew V. Frane
University of California Los Angeles
Los Angeles, CA

An SPSS script previously presented in this journal contained nontrivial flaws. The script should not be used as written. A call is renewed for validation of new software.

Keywords: effect size, confidence intervals, SPSS, syntax

Letter to the Editor

Walker (2015) presented an SPSS program for estimating effect sizes and approximating confidence intervals. It contains flaws and should not be used. The consequences are nontrivial, as is apparent from Walker's example, which used the following input: $M_1 = 16.45$, $M_2 = 11.77$, $SD_1 = 2.23$, $SD_2 = 4.66$, $N_1 = 30$, $N_2 = 34$, $N = 64$, where M_1 and M_2 are the sample means, SD_1 and SD_2 are the sample standard deviations, N_1 and N_2 are the group sample sizes, and N is the total sample size. Given this input, the resulting 95% confidence intervals in Walker's output (see his Table 1) are far too narrow: either [1.109, 1.403] or [1.094, 1.387], depending on whether Cohen's *d* or an approximation of Hedges' *g* is used in the estimation.

Walker did not validate these results by simulation, or by analytic methods, or by comparing the results to those produced by established software. For example, the `ci.smd` function in the extensively vetted MBESS package for *R* (see Kelley, 2007; Kelley & Rausch, 2006) uses a standard iterative procedure to compute exact confidence intervals for the standardized effect size. For Walker's input, the `ci.smd` function may be executed in conjunction with the `smd` function, as follows:

```
library (MBESS)
```

Mr. Frane is a doctoral student of cognitive psychology in the Multisensory Perception Lab. Email at avfrane@gmail.com.

```
cohend <- smd (Mean.1=16.45, Mean.2=11.77, s.1=2.23, s.2=4.66,
   n.1=30, n.2=34)
ci.smd (smd=cohend, n.1=30, n.2=34, conf.level=.95)
```

This method correctly gives the 95% confidence interval as [0.714, 1.790]. Note that this interval is much wider than Walker's approximations and is appropriately asymmetrical around Cohen's *d*.

Part of the problem with Walker's code is how it computes the variables it calls D1 and G1. These cryptically-named variables purportedly estimate the error terms of Cohen's *d* and Hedges' *g* (respectively), but as coded actually estimate the squares of those error terms. That is, the program computes estimated variances when it should be computing estimated standard errors. The same confusion is evident in Walker's equation 9 (compare to Hedges & Olkin, 1985, p. 86, equation 15, which appropriately squares the error term on the left side of the equation). Hence, Walker's erroneous computations could be vastly improved by adding square roots to the two lines of code where D1 and G1 are computed, as follows:

```
COMPUTE D1 = SQRT (N / (N1*N2) + COHEND**2 / (2*N)).
COMPUTE G1 = SQRT (N / (N1*N2) + HEDGESG**2 / (2*N)).
```

However, there is no justification for using approximations at all, given that superior, exact confidence intervals can now be easily computed with simple commands in freely available, industry standard software (namely, *R* with the MBESS package).

Walker acknowledged that by disregarding noncentrality, the program could not provide exact confidence intervals, a limitation defended as follows: "Bird (2002) found that if d is < 2.00, which in social science research frequently can be the circumstance with middling-sized effects (Richard, Bond, & Stokes-Zoota, 2003; Rosnow & Rosenthal, 2003), adjustment for noncentrality is not compulsory" (Walker, 2015, p. 285). Bird (2002) did note that heuristically speaking, approximate standardized intervals are likely to be similar to exact standardized intervals for $d < 2$, provided degrees of freedom ≥ 30. However, Walker overlooked Bird's caveat that "exact standardized intervals should be preferred to approximate standardized intervals whenever both are available" (Bird, 2002, p. 204).

Walker's program implements incorrectly a method that would be obsolete even if implemented correctly. The program also contains other peculiarities. For

example, given that the user must input N_1 and N_2, it is redundant that the program also requires the user to input N (which the program could instead have computed for itself, as simply $N_1 + N_2$). Additionally, an anonymous reviewer of the present letter identified a potentially confusing conflict between the coding and the text in Walker's article: The coding computes Cohen's *d* using the pooled standard deviation, which is likely the proper approach, but Walker's equation 6 computes Cohen's *d* using the unweighted average of SD_1 and SD_2.

Walker (2015) appeared in the same issue as an article noting the perils of using inadequately vetted statistical software (Lorenz, Markman, & Sawilowsky, 2015). Indeed, checking new software against established software prior to dissemination and professional use is essential.

References

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62*(2), 197-226. doi: 10.1177/0013164402062002001

Hedges, L. N., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8), 1-24. doi: 10.18637/jss.v020.i08

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*(4), 363-385. doi: 10.1037/1082-989X.11.4.363

Lorenz, A. J., Markman, B. S., & Sawilowsky, S. (2015). Caution for software use of new statistical methods (R). *Journal of Modern Applied Statistical Methods, 14*(2), 275-281. Retrieved from http://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=2021&context=jmasm

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363. doi: 10.1037/1089-2680.7.4.331

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57*(3), 221-237. doi: 10.1037/h0087427

Walker, D. A. (2015). Two group program for Cohen's *d*, Hedges' *g*, η^2, R_{adj}^2, ω^2, ε^2, confidence intervals, and power. *Journal of Modern Applied Statistical Methods, 14*(2), 282-292. Retrieved from http://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=2086&context=jmasm

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 783-784.
doi: 10.22237/jmasm/1493599380

ISSN 1538 − 9472

In Response to Frane

David A. Walker
Northern Illinois University
DeKalb, IL

A rebuttal to Frane's letter to the Editor in this issue.

Letter to the Editor

Frane (2017) disagreed with my interpretation of Bird (2002), suggesting I overlooked Bird's (2002) important assertion that "exact standardized intervals should be preferred to approximate standardized intervals whenever both are available" (p. 204). The ensuing sentence from Bird (2002) should be stated, because it is effective for full contextual purposes: "It is often necessary, therefore, to rely on approximate (classic) intervals for inferences about standardized effect sizes" (p. 204). A personal, research perspective is important, as is taking stock in this assertion from Bird (2002), which utilized, "In general, approximate and exact standardized intervals are likely to lead to similar (often indistinguishable) interpretations of effect sizes (p. 204)." Frane (2017) suggested the entirety of this idea was qualified under the pretext of heuristically speaking, but it is not clear how this could be known.

To be sure, there was full comprehension of Bird (2002), but exact confidence intervals (CIs) were not the intent of Walker (2015). This was obvious even with Frane's (2017) example and *R* code, because it "uses a standard iterative procedure to compute 'exact' confidence intervals for the standardized effect size". However, one of the main objectives, stated in the first sentence of Walker (2015), was to afford code in SPSS, not *R*. Moreover, as stated in Walker (2015) at numerous locations and with support from literature, "The program's estimated CI formula is based on previous research." The operative word was estimated and similar synonyms, such as approximate, but not, as Frane (2017) would have it, "exact."

Dr. Walker is a a Professor of Educational Research, Technology, and Assessment.
Email at dawalker@niu.edu.

Wrangling about the peculiarities of a program that a user might not be advocating in favor of alternative programs remains a personal choice. It should not, however, rise to a level warranting description as a fundamental flaw, obsolete, or an incorrect implementation. Frane (2017) claimed the program in Walker (2015) does not provide exact confidence intervals. Precisely. Exact CIs were not discussed in Walker (2015), because they did not comport with the purpose of the article.

References

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62*(2), 197-226. doi: 10.1177/0013164402062002001

Frane, A. V. (2017). Errors in a program for approximating confidence intervals (Letter to the Editor). *Journal of Modern Applied Statistical Methods, 16*(1), pp.-pp. doi: 10.22237/jmasm/1493599320

Walker, D. A. (2015). Two group program for Cohen's *d*, Hedges' *g*, η^2, R_{adj}^2, ω^2, ε^2, confidence intervals, and power. *Journal of Modern Applied Statistical Methods, 14*(2), 282-292. Retrieved from http://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=2086&context=jmasm

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 785-786
doi: 10.22237/jmasm/1493599500

ISSN 1538 − 9472

Instructions for Authors

Authors wishing to submit to *JMASM* may do so using the submission form at the journal's website, http://digitalcommons.wayne.edu/jmasm. Three areas are appropriate for *JMASM*:

1. Development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods;
2. Development or study of nonparametric, robust, permutation, exact, and approximate randomization methods; and
3. Applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Work appearing in *Regular Articles*, *Brief Reports*, and *Emerging Scholars* is externally peer reviewed, with input from the Editorial Board; work appearing in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* is internally reviewed by the Editorial Board. *JMASM* charges neither article processing fees nor submission fees.

Please observe the following guidelines when preparing manuscripts:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Articles should be submitted without a title page or abstract. There should be no material identifying authorship except in the fields of the submission form. Include a statement in the cover letter indicating that proper human subjects protocols were followed where applicable, including informed consent.
3. Manuscripts should be prepared in Microsoft Word (.doc or .docx) only (Wordperfect and .rtf formats may be acceptable − please inquire). Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are NOT acceptable for manuscript submission.
4. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
5. Create tables without boxes or vertical lines. Place tables, figures, and graphs "in-line", not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
6. The submission form requires an Abstract with a 50 word maximum, and a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and

References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left justified, indent optional.

7. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.

8. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use "&" instead of "and" in multiple author listings.

9. Suggestions for style: Instead of "I drew a sample of 40" write "A sample of 40 was selected". Use "although" instead of "while," unless the meaning is "at the same time." Use "because" instead of "since," unless the meaning is "after." Instead of "Smith (1990) notes" write "Smith (1990) noted." Do not strike the spacebar twice after a period.

Journal of
Modern Applied Statistical Methods

ISSN: 1538−9472 http://digitalcommons.wayne.edu/jmasm

PUBLISHED biannually (May, November) in partnership by:

JMASM, Inc.
PO Box 48023
Oak Park, MI 48237
ea@jmasm.com

Wayne State University Library System
Purdy Library
Detroit, MI 48202
digitalcommons@wayne.edu

To Advertisers

Advertisements are accepted at the discretion of the editor. Send requests for advertising information to ea@jmasm.com.

WAYNE STATE UNIVERSITY
LIBRARY SYSTEM